Q&A 커뮤니티 기반 전문영역 검색을 위한 프레임워크

A Framework for Q&A Community based Vertical Search

정옥란(Ok-Ran Jeong)*, 오제환(Jehwan Oh)**, 이은석(Eunseok Lee)**

초 록

본 연구는 Q&A(question and answer : 질문-답변) 커뮤니티 사이트에서 집단지성의 특성을 추출하고, 이를 이용한 전문지식이나 정보 검색을 위한 전문영역 검색(vertical search)을 위한 프레임워크를 제안한다. 많은 Q&A 사이트로부터 얻은 정보는 하나의 집단지성의 형태로 볼 수 있으며, 전문영역 검색은 특정 전문 분야 검색에 초점을 맞춘 검색 방법이다. 제안된 프레임워크는 사용자가 검색하고자 하는 질의어와 연관되어 있는 질문(question)과 답변(answer) 정보를 이용하여 관련어를 확장한 후, 이를 기반으로 전문지식을 요구하는 특정 도메인분야에 적용하게 된다. 이를 통해 일반 검색 엔진을 통해 검색된 검색 결과보다 유용한 정보와 전문적인 상세정보까지 제공해 줄 수 있다.

ABSTRACT

This study suggests a framework which extracts features of collective intelligence from social Q&A community sites and takes advantage of those features upon vertical search for domain specific knowledge or information retrieval. One source of collective intelligence on the internet is the question and answer(Q&A) data available from many Q&A sites. Vertical search is focused on searching special areas or specific domains. This paper proposes a framework for extending the relevant terms by using Q&A information connected with query that the user wants to retrieve, and then applies them to specific domain field that requires professional and detailed knowledge.

키워드 : Q&A 커뮤니티, 집단지성, 검색엔진, 전문영역 검색
Q&A Community, Collective Intelligence, Search Engine, Vertical Search

---

이 논문은 2010년도 경원대학교 교내연구비 지원에 의한 결과임.
* 교신저자, 경원대학교 소프트웨어 설계・경영학과
** 성균관대학교 컴퓨터공학과
2011년 04월 08일 접수, 2011년 05월 16일 심사완료 후 2011년 05월 20일 게재확정.

1. 서 론

웹 기반 디지털 산업이 국내외 경제의 중요한 역할을 담당하며 차세대 기술 환경에서 소셜 네트워크 기반 기술이 급속도로 발전하고 있다. 본 연구에서는 새로운 패러다임인 소셜 네트워크(social network)환경에서 상호 질문과 답변을 하는 커뮤니티 기반 지식검색 정보를 이용하며, 이 정보를 전문영역 검색(vertical search)에 적용하기 위한 프레임워크를 제안하고자 한다. 전문영역 검색 엔진은 모든 도메인을 대상으로 하는 것이 아니라 특정 전문 분야 또는 도메인 내에서만 검색하는 것을 말한다. 이러한 검색은 현존 검색을 이용했을 때 대량의 결과를 보여주고 있지만, 전문적인 지식이나 정보를 사용자가 원할 경우 해당 도메인 분야로 범위를 좁혀 초점을 맞추기는 어려운 현실이다.

전형적인 정보 검색은 디지털 라이브러리 및 온라인 검색 사이트에서 대부분의 지식 및 정보 검색들이 이뤄지고 있다. 소셜 네트워크 기반 온라인 Q&A 사이트[7]가 활성화 되면서 사용자들의 검색에 대한 요구는 갈수록 다양해지고 정확하고 만족스런 답변을 기대하고 있다. 일반적인 검색의 급속한 발달로 찾고자 하는 정보에 대한 검색 결과를 대량으로 얻을 수 있다. 하지만, 어떤 특정 전문영역에서의 지식검색은 쉽지 않은 것이 현실이다.

다음 간단한 질문에 대해서 생각해보자. “알츠하이머병”에 대한 검색을 하고자 한다면, 사용자는 이병에 대한 발병 원인, 이병을 경험한 사람들의 다양한 답변 또는 치료 및 관리를 위한 전문 지식까지도 알고자 할 것이다. 단순한 검색으로는 이 질문에 대한 답을 해 줄 수 없는 상황이다. 사용자는 지식 검색 서비스를 해주는 포털 사이트인 ‘네이버 지식iN’[1], ‘Yahoo! Answers’[23], ‘Twitter’[18], ‘Orkut’ 등에 질의를 올려 답변을 얻는 방법을 이용할 것이다. 위의 질문에서 얻은 답변들은 사용자 참여를 바탕으로 하는 사용자들 간의 상호작용에 초점을 맞춘 Q&A 기반 집단지성(collective intelligence)을 이용한 검색 방법이다. 이와 같은 방법은 많은 검색 대상 콘텐츠 중에 사용자가 원하는 가치 있는 정보를 찾기 위하여 집단지성을 이용한 것이다.

본 연구에서는 집단지성을 이용하기 위한 첫 단계로 질문-답변 커뮤니티를 분석한다. 질문-답변 커뮤니티는 사용자가 작성한 질문에 대한 다른 사용자들의 답변들을 수집 및 분석 한 후, 질문자가 답변들 중 가장 적절한 답변 들을 선택한다. 이러한 소셜 네트워크를 이용할 때, 해당되는 질문-답변의 도메인을 한정한다면, 다양한 소셜 요소를 담고 있는 검색 결과를 찾는데 효율적일 것이다.

다음 단계로는 전문영역 검색을 수행한다. 첫 단계에서의 소셜 요소를 담고 있는 검색 결과일지라도, 전문적인 지식을 요구하는 전문분야에 대한 검색은 일반검색으로는 찾기 어렵다. 예를 들면 의학관련, 법률관련, 학교나 연구소의 연구주제관련, 행정부서의 담당, 유사 기업 내 특정분야 담당 등에 대한 검색은 쉽지 않은 것이 현실이다. 정확하게 한 분야(도메인)를 겨냥하여 검색하고자 하는 주제를 찾아가는 것이 전문영역 검색(vertical search, domain specific) 기법이다. 이는 최상급의 주류 검색 엔진에서 대면하게 되는 수십억 개의 검색 결과와는 차별적인 것이다. 일반적인 검색이 아닌, 특별한 목표가 있는

검색이라면 전문영역 검색이 적합하며, 이를 이용한다면 관련된 분야의 보다 정확하고 수준 높은 검색 결과를 얻을 수 있다.

본 연구에서는 이러한 Q&A 커뮤니티 정보와 전문영역의 지식을 기반 한 검색 위한 프레임워크를 제안하고자 한다. 제안된 프레임워크는 다음과 같이 세 부분으로 구성되었다.

첫째, 사용자가 검색을 위해 검색어 또는 질의어를 주었을 때, 해당 문장의 의미와 도메인을 판단하기 위한 키워드 분석 알고리즘을 적용한다. 여기에서 분석을 위한 전처리가 이루어진다(전처리 모듈(preprocessing module)).

둘째, 키워드 분석을 통해 어떤 도메인의 검색이 이루어져야 하는지 결정된 후 그 분야의 소셜 네트워크상에 있는 사용자들의 질의-답변 정보를 분석한다. 유용한 관련 정보를 확장하기 위해 질문-답변 집합을 이용하여 해당 카테고리를 확장하고, 그 확장된 카테고리에 따른 핵심어들을 찾는다(질문-답변 모듈(Q&A module)).

셋째, 확장된 카테고리에 따른 핵심 키워드를 이용하여 매칭되는 분야별 전문영역 검색 엔진을 연결시켜준다. 매칭된 전문검색 엔진이 해당 분야에 따른 전문영역 검색을 수행한다(전문영역 검색 모듈(vertical search module)).

본 논문의 구성은 다음과 같다. 제 2장에서는 Q&A 시스템과 전문영역 검색에 대한 관련연구를 설명하고, 제 3장에서는 본 연구에서 제안하는 프레임워크에 대해 전체적인 구성에 대해 설명한다. 제 4장에서는 제안한 프레임워크를 단계별로 설명하고, 단계별 데이터를 적용하는 과정을 기술한다. 제 5장에서는 실험 및 평가를 하며, 제 6장에서 결론을 서술한다.

2. 관련 연구

'네이버 지식iN', 'Yahoo! Answers'와 같이 지식 검색을 위한 질문-답변 커뮤니티는 사용자가 작성한 질문에 대해 다른 사용자가 답변을 하는 게시판 형태의 공유 서비스이다. 질문-답변 커뮤니티는 의견, 조언, 노하우 등과 같이 일반적인 검색엔진으로 쉽게 찾을 수 없는 형태의 정보도 질문을 작성하면 다수로부터 답변을 얻을 수 있다. 이러한 장점으로 정보를 얻는 새로운 창구로서 지난 몇 년 간 지식검색 서비스로 주목받고 있다.

지식검색 서비스의 형태는 온라인 게시판이나 유즈넷과 유사한 형식으로 이루어졌다. Zhongbao[9]는 온라인 게시판을 이용한 소셜 네트워크 분석을 수행하여 사용자들의 행동 패턴이 그들의 관심 공간에 따라 달라지는 것을 보였다. Jeon[8]은 답변의 유사도를 통해 의미적으로 비슷한 질문을 구한 후, 이들에 관련 있는 키워드들을 추출하여 검색의 성능을 향상 시켰다. Agichtein[5]은 질문-답변 문서의 길이, 구두점의 수와 같은 텍스트 정보뿐만 아니라 답변자의 등급, 추천수, 조회수 등의 비텍스트 정보들을 사용하여 개선된 문서의 품질 측정 방법을 제안하였다. Whittaker[13]은 유즈넷을 대상으로 사용자의 수, 글의 길이 등의 통계적인 패턴을 찾아내었다. Adamic[10]은 'Yahoo! Answer'의 카테고리들

을 K-means 알고리즘을 사용하여 문서의 길이, 질문당 답변의 개수, 사용자간의 상호작용 패턴 등의 특성에 따라 세 가지 분류로 클러스터링 하였다.

Aardvark 회사는 소셜 네트워크 기반 검색 엔진인 'Aardvark'를 개발하였다[4]. 사용자가 질문을 인스턴트 메세지(instant message), 이메일, 웹 입력, 텍스트 메시지(text message), 아이폰(iPhone)을 이용한 음성 입력, 'Twitter' 등으로 하게 되면, 사용자의 확장된 소셜 네트워크 내에서 해당 질문에 답할 수 있는 사람을 찾고, 그에 대한 답변을 찾는 형식이다. 이 엔진은 답변을 라이브러리에서 찾을 뿐만 아니라, 소셜 네트워크 기반 빌리지 패러다임(village paradigm)을 이용하여 사용자의 친밀감을 기반으로 하였다. Aardvark는 게이트웨이를 통해 질문이 들어오면 질문을 분석하여, 중앙에 대화 매니저(conversation manager)가 사용자의 관련도를 데이터베이스에서 찾는다. 이후 확장된 소셜 네트워크를 이용하여 해당 질문에 대해 답변할 수 있는 사람들을 검색하여, 그 사람의 정보와 일반적인 검색결과를 함께 질문자에게 제공하는 것이다.

사용자의 요구에 맞게 웹 문서를 제공하기 위한 노력은 많이 이루어져왔다. [2]에서는 개념 망을 이용하여 사용자의 요구와 문서간의 개념 매칭도 계산을 통하여 개념적으로 관련성이 높은 문서를 제공하였다. 또한, 사용자가 필요로 하는 정보만을 정확하게 검색하기 위해 한 분야만을 겨냥하여 만들어진 전문영역 검색 엔진이 개발되고 있다. 전문영역 검색은 한 분야만의 전문적인 검색결과를 보여주는 대표적인 전문영역 검색 사이트는 'SearchMedica'[16], 'WebMD'[21] 등이 있다. 사용자가 해당 사이트에 접속하여 병명을 검색하게 되면, 이에 관련된 최근 의학 연구 기사와 논문, 치료 방법, 환자가 알아둘 내용, 치료가 될 만한 식품, 운동 및 의약품 등의 정보를 카테고리별로 나누어 상세한 검색 결과를 보여준다. 여러 가지 전문영역 검색 엔진을 카테고리별로 모아 제공하는 멀티 전문영역 검색 사이트들은 'Eurekster'[6], 'Search engine guide'[14], 'Virtual sites'[20], 'Search engine index'[15] 등이 있다. 전문영역 검색 포털 사이트는 질문-답변 커뮤니티에 대한 검색을 '지식검색'이라는 서비스를 통해 제공하고 있다. 하지만 일반적인 웹 문서와 사용자에 의해 생성된 콘텐츠와의 상이성과 질문-답변 콘텐츠의 특수성 등으로 인해 발생하는 문제점을 해결하기 위한 많은 연구가 이루어지고 있으나 높은 신뢰도의 검색 결과 제공은 아직 미흡한 설정이다.

질문-답변 커뮤니티는 사용자의 질문에 대해 다른 사용자들이 답변을 하고, 많은 사용자들이 이 지식의 유용성에 대해 평가함으로 지식을 공유한다. 이러한 방법은 사용자 다수가 인정하면 유용한 답변으로 채택하게 되어 있다. 이는 질문자가 답변을 주관적으로 선택하게 되면 그것이 정답이 아니라도 정답인 것처럼 왜곡될 수 있고, 방대한 양의 질문과 답변 중에서 불필요한 정보와 출처가 불분명한 정보들로 인해 지식검색의 신뢰도가 낮아질 수 있기 때문이다. 사용자의 의도에 맞는 정확한 지식 서비스를 제공하기 위해 사용자의 질문에 대한 정확한 분석이 필요하며, 해당 질문에 작성된 주관적인 답변 이외에도 소셜 네트워크에 존재하는 답변으로 사용 가능한 프로파일(profile)과 히스토리(history)를

검색하여 객관적인 답변으로 제공해야 한다. 이 과정에서 사용자의 질문에 대한 분석이 차별화되어 정확하게 이뤄져야 할 것이며, 이 분석 후 답변을 얻을 수 있는 기반 기술이 추가적으로 따라야 할 것이다.

소셜 검색을 이용한 Aardvark 시스템도 질문에 대한 분석 후 전체 도메인을 대상으로 답변을 찾고 있다. 그러나 전문지식을 요구한다면, 전문적인 지식 정보에 대해서는 한계가 나타나게 되어 있다. 이러한 점을 보완하기 위해 본 연구에서 제안한 Q&A 커뮤니티 기반 전문영역 검색을 위한 프레임워크를 이용한다면, 아직까지 얻기 어려웠던 전문영역의 결과를 얻을 수 있을 것이다. 예를 들어, 의료 관련 질의에 대해서, Q&A 사이트들에서 관련 정보와 키워드들을 얻어, 검색 키워드들을 관련 키워드들로 확장하여, 전문영역 검색을 하게 되는 것이다. 확장된 키워드를 이용하여 전문영역 검색을 하게 된다면 관련된 현재까지의 연구결과, 환자에게 필요한 정보, 관련 치료 병원에 대한 정보, 증명이 된 최근 의학 기사, 논문 등 전문 정보들을 찾을 수 있게 되는 것이다.

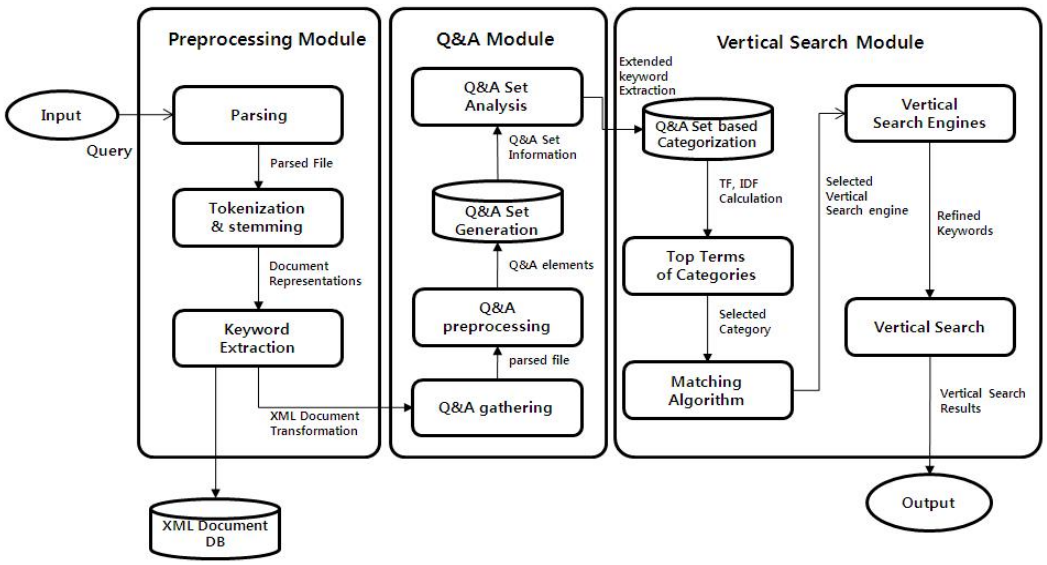
3. 제안 프레임워크

소셜 네트워크를 기반으로 생성된 소셜 웹 사이트들은 질문-답변 정보를 제공하는 서비스를 가지고 있다. 이러한 질문-답변 정보 활용을 통해 보다 유용한 집단지성의 장점을 활용 가능하다. 본 논문은 기존 연구[22]에서 제안한 소셜 웹 사이트들의 주요 특성(essential feature)들을 이용할 수 있다. 일반적인 키워드 검색이 아닌 전문분야나 제한된 특정 분야에서의 검색을 하고자 할 때 그 해당 분야에 맞는 전문영역 검색이 이뤄져야 할 것이다.

본 연구에서는 검색어나 질의어에서 추출된 키워드의 의미를 확장하기 위하여 질문-답변 기반 소셜 네트워크를 이용한다. 이를 통해 해당 키워드로 확장 가능한 도메인과 핵심어에 대한 정보를 획득하여, 해당 도메인에 맞는 전문영역 검색 엔진을 이용한다. 즉 본 연구에서는 해당 분야의 소셜 네트워크상의 질문-답변 정보를 활용하고, 정확한 분석에 따른 전문영역 검색과 이원화된 소셜 네트워크를 이용한다. 질문 분석 단계에서 정확한 내용 파악 및 해당 도메인을 알아 낼 수 있다면 전문영역 검색 엔진 기법을 이용해서 정확하고, 전문적인 답변을 얻을 수 있을 것이다.

<그림 1>은 본 논문에서 제안하는 전문영역 검색 시스템의 프레임워크의 전체 구조도를 보여준다. 제안 프레임워크는 세 개의 모듈로 구성된다. 기본적인 작업을 위한 전처리 모듈(preprocessing module), 질문-답변 내용을 확장하기 위한 질문-답변 모듈(Q&A module), 전문영역 검색 엔진에 적용해서 결과를 산출해내는 전문영역 검색 모듈(vertical search module)로 구성된다.

검색하고자 하는 사용자 질의가 입력되면, 해당 질의를 분석 한 후 그에 해당되는 사이트들의 질문-답변을 분석하여 해당 질의에 맞는 확장된 검색 카테고리를 결정한다. 결정된 카테고리를 기반으로 가장 필요로 하는 전문영역 검색 서비스 분야를 찾는다. 이 결과에 매칭되는 전문영역 검색 엔진의 매칭 알고리즘 적용한다. 매칭된 해당분야의 전문영역 검색엔진을 통해 검색결과를 보여준다.



〈그림 1〉 제안 프레임워크의 전체 구조

3.1 전처리 모듈(Preprocessing Module)

특성 추출 및 Q&A 모듈에 적용하기 위한 데이터 전처리 알고리즘 적용한다. 연구 단계에서 자연어 처리가 포함될 경우 데이터의 전처리 알고리즘은 필수적으로 적용되어야 한다. 특히 소셜 네트워크 사이트에 존재하는 정보는 UCC(user created contents)나 웹 게시판 등과 같이 공식적인 포맷에 맞추어 작성되지 않는다. 전처리 과정은 어근화 및 불용어 제거뿐만 아니라 사용자들이 자주 사용하는 약어나 은어에 대한 처리도 고려되어야 한다. 문장 분석을 위해서는 먼저 질의어(query)가 입력된 후 검색된 일차적인 결과 문서를 파싱(parsing)하고, 토큰으로 나누는 작업(tokenization)이 이뤄져야 한다. 불용어 제거와 어근화 처리(stemming)를 하여 주요 키워드들만을 추출한다.

3.2 질문-답변 분석 모듈(Q&A Module)

추출된 주요 키워드들을 이용하여 Q&A 커뮤니티 사이트에서 관련된 질문-답변 정보들을 찾아낸다. 이를 통해 관련된 검색 키워드 외에 이에 관련하여 유용한 사용자들의 정보들을 얻을 수 있을 것이다. 얻어진 정보를 활용하기 위해 특성 중 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 TF(term frequency : 단어빈도)와 특정한 단어가 대상으로 하는 전체 문서 중에 몇 개의 문서에서 출현하는 지를 나타내는 DF(document frequency : 문서빈도)를 추출한다. 이는 추출된 용어들에게 그 중요도에 따라 가중치를 주기 위함이다. 세부 단계는 먼저, 질문-답변 정보를 수집하는 질문-답변 자료를 모으고, 이 데이터에 대한 전처리를 수행 한다. 다음 단계로는 질문-답변 분석 단계로 질의 키워드

로 검색된 질문-답변 집합을 이용하여 TF, IDF(inverse document frequency)를 계산하여 질문-답변 집합에 등장하는 각 용어의 중요도를 알아낸다.

3.3 전문영역 검색 모듈(Vertical Search Module)

전문영역 검색 모듈에서는 전 단계 결과물로 나온 질의어와 관련된 질문-답변 키워드에 해당되는 전문영역 검색 엔진들을 매칭 시킨다. 키워드를 중심으로 전문영역 검색 엔진을 실시간으로 크롤링(crawling) 할 수도 있고, ODP (open directory project)[12] 항목에 따라 기존의 전문영역 검색 엔진을 이용할 수 있다. 본 연구에서는 ODP 기반 목록별로 기존의 전문영역 검색엔진을 매칭하여 결과를 보여준다.

4. 제안 프레임워크 구현

본 연구에서는 기존의 검색 엔진에 질문-답변에 대한 정보를 이용하고, 이 정보를 전문검색 엔진에 적용하는 프레임워크를 제안하였다. 제안한 프레임워크를 설계 및 구현하여 적용하는 과정을 단계별로 설명한다. 첫 단계에서는 프레임워크 전처리 및 분석 모듈을 세부단계별로 보여주고, 두 번째 단계에서는 전문영역 검색 엔진 적용 방법 및 결과를 보여준다.

4.1 프레임워크 전처리 및 분석 모듈

4.1.1 질의어 전처리

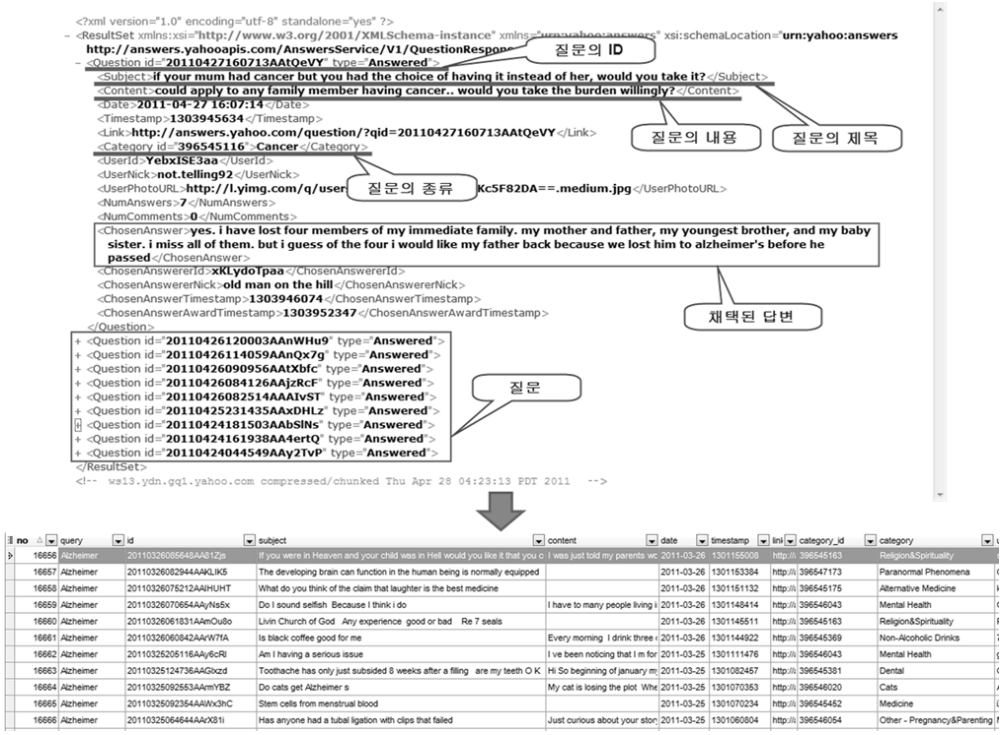
사용자 질의어를 분석하기 위하여 처음 파싱을 하는데, 파싱 단계에서는 사용자 질의어를 토큰으로 나눈다. 그 다음단계로는 각 토큰의 불용어 제거와 어근화 처리를 통한 키워드 추출 단계이다. 불용어를 제거하기 위해서는 각 용어의 속성 파악이 필요하다. 우리는 미국 카네기 멜론 대학교(Carnegie Mellon University)에서 수행한 'Bracketing Guidelines for Treebank II Style Penn Treebank Project'[3]에서 사용한 용어 태깅 방법을 이용하였다. 이 프로젝트에서는 7만여 개의 용어에 대한 라이브러리를 제공한다. 이를 이용하면 사용자 질의에 출현하는 토큰들의 속성 파악이 가능하다. 어근(stem)이란 단어에서 접두사와 접미사를 제거하고 남는 부분이다. 보통 사용자에 의해 질의 된 단어와 검색 시의 관련 문서들에 포함된 단어가 실제 내용은 일치하지만, 구문적 변형으로 인해 검색 결과의 성능을 저하시키는 원인이 될 수 있다. 이러한 문제는 각 단어를 어근으로 대체하면 해결 할 수 있다. 본 연구에서는 어근화 처리를 하기 위해서 'The Porter Stemming Algorithm' [17]을 이용하여 각 토큰에 대한 어근을 찾았다. 'The Porter Stemming Algorithm'은 6~7단계를 거치면서 37개의 룰을 적용하여 용어의 어근을 찾아주는 대표적인 알고리즘이다. 마지막 단계에서는 불용어 제거와 어근화 처리를 거친 후에는 검색을 위하 토큰을 선정한다. 우리는 검색을 위하여 키워드로 명사를 형태소로 가진 키워드를 선정하였다.

4.1.2 질문-답변 정보 수집

이번 단계에서 질문-답변 정보 수집을 위하여 'Yahoo! Answers' 사이트를 이용하였다. 'Yahoo! Answers' 사이트에서 제공하는 API를 통해 사용자 질의 전처리를 통해 추출한

각 키워드에 대한 검색 결과를 XML 형식으로 얻을 수 있다. <그림 3>은 XML 형식으로 수집한 질문-답변 문서를 보여준다. 질문-답변 문서는 'question id', 'subject', 'content', 'chosen answer', 'category id' 등의 정보 포함한다.

〈그림 2〉 전처리 단계를 통해 키워드 추출 : "Alzheimer's Disease"라는 질의를 통해 Alzheimer와 Disease라는 키워드를 추출



〈그림 3〉 수집된 XML 문서가 데이터베이스에 저장 : XML 문서는 'question id', 'subject', 'content', 'chosen answer', 'category id' 등의 정보 포함

4.1.3 질문-답변 집합 생성

질문-답변 문서는 일반 문서와 달리 기본적으로 질문과 답변 한 쌍으로 존재하며, 다양한 정보들이 추가적으로 포함되어 있다. 우리는 질문-답변의 이러한 특성을 표현하기 위해서 질문-답변을 하나의 집합으로 구성하였다. 질문-답변 집합은 question title(Q_1), question contents(Q_2), best answer(A_1), other answers(A_2)로 구성된다. 질문-답변 집합은 식 (1)과 같이 표현된다.

$$Q \ \& \ A\_Set = \{Q_1, Q_2, A_1, A_2\} \quad (1)$$

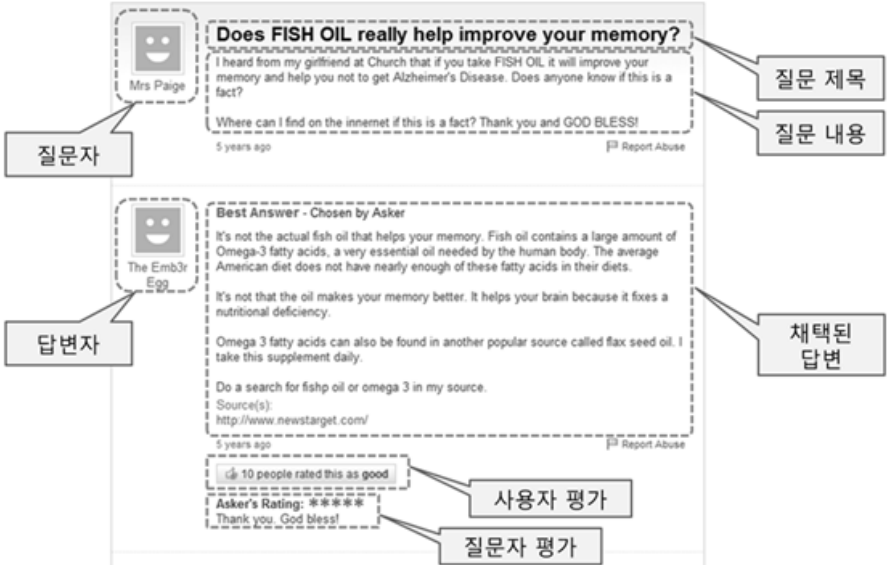
Q_1 : 질문의 제목
Q_2 : 질문의 내용
A_1 : 질문자가 채택한 답변
A_2 : A_1을 제외한 답변

그리고 질문-답변 집합을 이루는 원소들은 식 (2)와 같이 표현된다.

$$\begin{aligned} Q_1 &= \{t_{q1_1}, t_{q1_2}, t_{q1_3}, \cdots, t_{q1_{n'}}\} \\ Q_2 &= \{t_{q2_1}, t_{q2_2}, t_{q2_3}, \cdots, t_{q2_{n''}}\} \\ A_1 &= \{t_{a1_1}, t_{a1_2}, t_{a1_3}, \cdots, t_{a1_{n'''}}\} \\ A_2 &= \{t_{a2_1}, t_{a2_2}, t_{a2_3}, \cdots, t_{a2_{n''''}}\} \end{aligned} \quad (2)$$

위의 식 (1)과 식 (2)는 <그림 4>에서 볼 수 있는 질문자의 질문 제목, 질문 내용, 채택된 답변과 그 의외의 답변들을 표현해주는 것이다. 또한 질문과 답변의 내용이 식 (1)과 식 (2)의 내용으로 <그림 3>의 'XML Document DB'에 저장된다.

<그림 4>는 'Yahoo! Answers' 사이트에서 제공하는 질문-답변의 구성을 보여준다. 이 사이트에서도 기본적으로 'question title', 'question contents', 'answer contents', 'best answer'를 제공하고, 추가적으로 'asker', 'user's



〈그림 4〉 Yahoo! Answer에서 제공하는 질문-답변 화면

rating', 'date' 등의 부가적인 정보를 제공한다.

4.1.4 질문-답변 전처리

생성된 질문-답변 집합의 의미를 분석하기 위해서는 전처리 과정이 필요하다. 먼저, 질문-답변 전처리 과정은 질의어 전처리와 같은 과정(파싱, 불용어 제거, 어근화 처리)을 수행하여 핵심 키워드들을 선정한다.

4.1.5 질문-답변 분석

질문-답변 분석 단계에서는 질의 키워드로 검색된 질문-답변 집합을 이용하여 TF, IDF를 계산하는 단계이다. 우선 같은 카테고리를 가지는 질문-답변 집합을 분류하고, 어떤 용어가 특정 질문-답변 집합에서 얼마나 중요한 것인지를 나타내기 위하여 TF, DF를 계산한다.

4.2 전문영역 검색 모듈(Vertical Search Module)

전문영역 검색 모듈에서는 질의어와 관련된 질문-답변 핵심 키워드들을 이용하여 해당 전문영역 검색 엔진을 매칭 한 후, 매칭 된 검색 엔진을 이용하여 재검색을 한다. 전문영역 검색 엔진의 매칭은 질문-답변 핵심 키워드를 분석하여, 각 해당 분야의 전문영역 검색엔진선택하게 된다. 선택된 전문영역 검색 엔진을 통해 해당 카테고리별 검색 결과를 보여준다. 분야 별 주요 전문영역 검색 엔진들은 <표 1>과 같다.

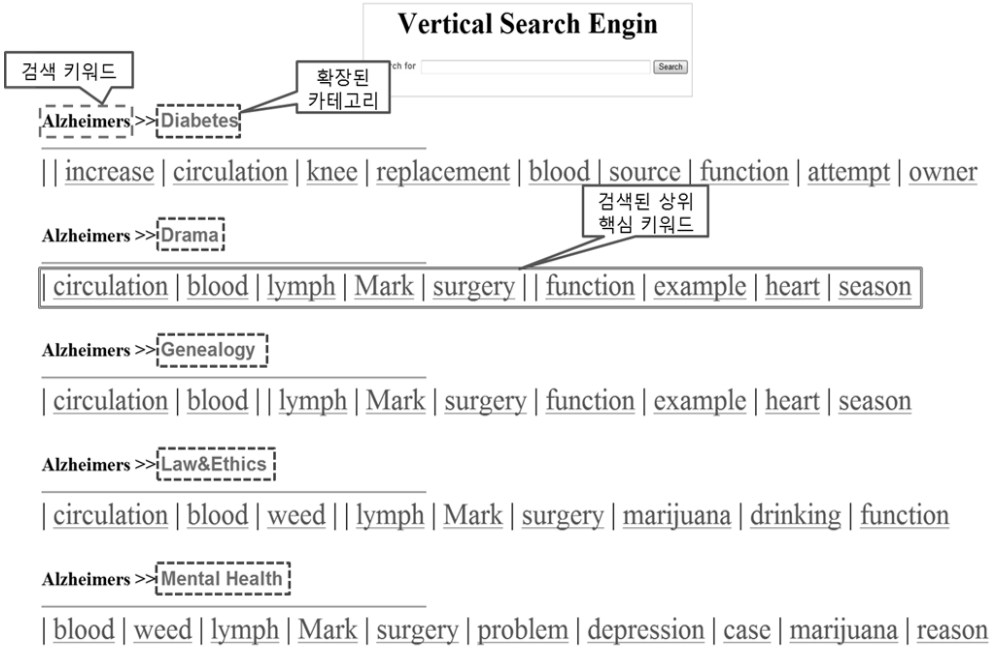
〈표 1〉 분야별 주요 전문영역 검색 엔진

분야	전문영역 검색 엔진
Health, Medicine	medstory, healthline, Google health, searchmedica, WebMD
Travel	kayak, TripAdvisor, Farecast
Events	Zvents, Eventful
Product [Business]	Amazon, theFind, Shopping.com
People [Society]	Spock
Music	Seeqpod, Grooveshark
Autos, Trips	Kosmix

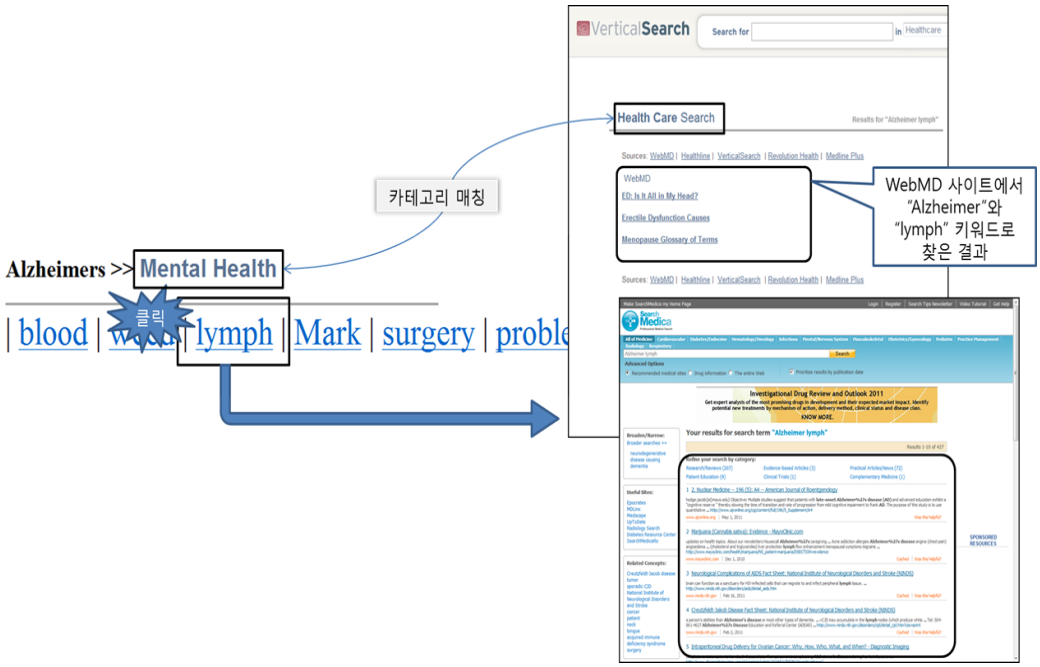
본 논문에서는 사용자가 질의한 키워드를 질문-답변 서비스를 통하여 카테고리 확장이 가능하다. <그림 7>은 "Alzheimer"라는 키워드를 제안 프레임워크에서 검색한 결과이다. 키워드 "Alzheimer"는 9개의 카테고리로 분류될 수 있다. 또한, 각 카테고리에는 핵심 키워드를 보여준다. 핵심 키워드는 질문-답변 분석 모듈 단계에서 계산된 TF-IDF를 이용하여 각 카테고리에서 가장 높은 7개의 용어를 나타낸다.

핵심 키워드를 클릭하면 사용자 질의 키워드와 핵심 키워드 그리고 카테고리를 기반으로 전문영역 검색을 시작한다. <그림 5>에서는 전문영역 검색 엔진으로 'Vertical Search [19]'가 선택되었으며, 제안 프레임워크를 통해 추출한 사용자 질의 키워드, 핵심 키워드, 카테고리를 이용하여 전문영역 검색 결과를 보여준다.

<그림 6>은 'Alzheimer' 키워드를 이용하여 검색된 7개의 카테고리 중에 'mental health' 카테고리에서 핵심어로 출현한 'lymph' 키워드를 선택하였을 때 'Vertical search' 사이트에서 보여주는 화면이다. 'Vertical search' 사이트에서는 'health care'와 관련된 5개의 전문영역 검색 엔진을 통해 검색된 결과를 보여주었다.



〈그림 5〉 사용자 질의 키워드에 대한 확장된 카테고리와 각 카테고리의 핵심 키워드



〈그림 6〉 각 핵심 키워드를 클릭하여 전문영역 검색을 통하여 보여주는 결과

5. 실험 및 평가

실험은 특정 질의어에 대해서 제안 시스템을 통해 추출한 핵심 키워드, 일반 검색 사이트에서 제공하는 검색 결과 페이지에서 추출된 TF-IDF 값이 높은 키워드 그리고, 전문 검색 영역 사이트에서 제공하는 검색 결과 페이지에서 추출된 TF-IDF 값이 높은 키워드를 비교하였다. 제안 시스템을 통해 추출한 핵심 키워드들과 일반 검색 사이트와 전문영역 검색 사이트에서 추출된 핵심 키워드들이 어느 정도 일치하는지를 실험한 것이다. 일치도가 높을수록 제안한 프레임워크가 유용하다고 판단할 수 있을 것이다.

5.1 실험 데이터

제안 시스템에서 핵심 키워드 추출을 위하여 사용한 데이터는 Yahoo! Answers에서 답변 작성이 완료된 질문들을 대상으로 수집하였다. Yahoo! Answers에서 질의어 'Alzheimer'를 검색하여 839개의 질문-답변 집합을 수집하였고, 이 데이터에서 15개의 카테고리가 발견되었다. 우리는 실험을 위하여 15개의 카테고리 중 3개의 카테고리(medicine, mental health, health)에 포함되어있는 질문-답변 집합 278개를 대상으로 하였다. 또한, 우리는 일반 검색을 위하여 해당 질의어에 대해여 Google과 Yahoo!에서 제공하는 검색결과 중 각각 상위 10개의 페이지에 대한 문서를 수집하였다. 그리고 전문 검색 영역은 Vertical Search 엔진에서 3개의 카테고리와 맵핑되는 카테고리 'health care'에서 제공하는 'WebMD' 사이트와 Healthline 사이트를 이용하여, 해당 질의어에 대한 검색결과 중 각각 상위 10개의 페이지에 대한 문서를 수집하였다. 실험을 위해 수집된 데이터는 <표 2>와 같다. 우리는 실험을 위해서 발견된 단어 중에 제 4장에서 기술한 전처리 단계를 수행하여 실험에 이용할 수 있는 단어를 추출하였다. 발견된 단어들 중에 실험에 사용된 단어의 비율은 평균 약 9.7%이다.

5.2 실험 결과

우리는 제 5.1절에서 언급한 데이터를 이용하여 제안 시스템, 일반 검색, 전문영역 검색에서 핵심 키워드를 추출하였다. 핵심 키워드 추출은 TF-IDF를 이용하였다. 제안 시스템에서 추출된 핵심어의 개수에 따라 일반 검색과 전문영역 검색에서 나타난 핵심 키워드

〈표 2〉 실험을 위해 수집한 데이터

구 분	수집 사이트	수집된 문서 개수	발견된 단어 개수	실험에 사용한 단어 개수	이용률(%)
제안 시스템	Yahoo! Answer	278	7,305	873	11.95
일반 검색	Google	10	11,188	845	7.55
	Yahoo	10	9,612	934	9.72
전문영역 검색	WebMD	10	3,741	398	10.64
	Healthline	10	3,133	276	8.81

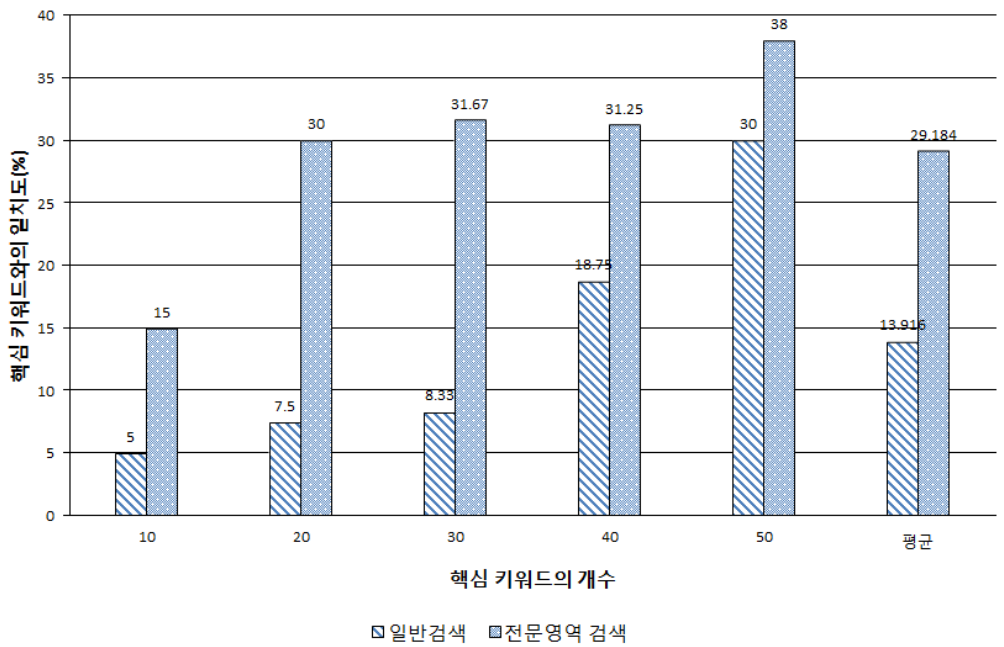
의 일치도를 계산하였다. <표 3>은 제안 시스템에서 추출된 핵심어의 개수에 따라 일반검색과 전문영역 검색에서 나타난 핵심어의 개수를 보여준다.

<그림 7>은 핵심 키워드 증가에 따른 일반검색과 전문영역검색의 핵심 키워드의 일치도를 보여준다. 전문영역 검색은 평균 약 29%의 일치도를 보였다. 하지만, 일반검색은 평균 약 13%의 일치도가 나타났다. 일반검색과 전문영역 검색에서 추출한 핵심 키워드의 개수가 증가하면서 일치도가 증가하는 것을 볼 수 있다. 이것은 핵심 키워드의 개수가 많아지면서 제안 시스템을 통해 추출한 핵심 키워드를 포함할 확률이 높아지기 때문이다. 핵

〈표 3〉 실험 결과

(단위 : 개)

구 분 \ 핵심 키워드 개수		10	20	30	40	50
일반 검색	Google	1	2	4	11	16
	Yahoo	0	1	1	4	14
	평균	0.5	1.5	2.5	7.5	15
전문 영역 검색	WebMD	2	7	12	16	20
	Helthline	1	5	7	9	18
	평균	1.5	6	9.5	12.5	19



〈그림 7〉 핵심 키워드 증가에 따른 일반검색과 전문영역 검색 핵심 키워드의 일치도

심 키워드 개수가 특정 수치를 넘어가면 일치도가 100%를 보일 것으로 예상한다. <그림 7>에서 보는 것과 같이 제안 시스템을 통하여 추출한 핵심 키워드가 일반 검색의 결과에서 보다 전문영역 검색에서의 결과에 더 많이 출현한 것을 확인할 수 있었다. 제안 시스템에서 추출한 핵심 키워드는 전문영역에서 추출한 키워드라고 볼 수 있다. 이것은 사용자가 제안 시스템을 통해서 전문영역 검색 결과를 바로 얻을 수 있음을 의미한다. 향후연구에서는 같은 양의 핵심 키워드에서 제안시스템에서 추출한 핵심 키워드와의 일치도를 높일 수 있는 방법이 필요할 것으로 보인다.

6. 결 론

본 연구에서는 Q&A 커뮤니티 기반으로 집단 지성의 정보와 특정 분야의 전문영역 정보를 동시에 얻을 수 있는 프레임워크를 제안하고 구현하였다. 본 연구에서는 프레임워크를 중점으로 설계 및 개발한 후, 실제 데이터를 가지고 단계별로 적용하였다. 제안한 프레임워크가 유용하다는 사실을 알 수 있었다. 향후 연구에서는 전문성 및 실용화를 고려하여 성능 분석을 위한 실험을 좀 더 진행할 것이다. 제안된 기술이 적용된 검색 엔진을 사용했을 때의 검색 결과는 사용자의 만족도를 높여줄 것이며, 검색 결과 자체의 전문 영역 지식 정보의 신뢰성을 향상 시켜줄 수 있을 것이다. 특히, 제안된 기술이 기업 내 엔터프라이즈 콘텐츠 관리 시스템에 적용된다면, 기업 내 기술이나 전문지식을 찾고자 할 때 유용하게 활용될 것이다.

참 고 문 헌

[1] 네이버 지식IN, http://kin.naver.com/.

[2] 이재원, 박성찬, 이상근, 박재휘, 김한준, 이상구, “개념 망을 통한 전자 카탈로그의 시맨틱 검색 및 추천”, 한국전자거래학회지, 제15권, 제3호, pp. 131-145, 2010. 8.

[3] Bracketing Guide lines for Treebank II Style Penn Tree bank Project, http://bulba.sdsu.edu/jeanette/thesis/PennTags.html/.

[4] Damon Horowitz, Sepandar D. Kamvar, “The Anatomy of a Large-Scale Social Search Engine,” Proc. of the WWW, pp. 4331-440, 2010.

[5] Agichtein, E., Castillo, C., Donato, D., Gionis, A. and Mishne, G., “High-quality content in social media,” Proc. of the International Conference on Web Search and Web Data Mining, pp. 183-194, 2008.

[6] Eurekster, http://www.eurekster.com/.

[7] Maxwell Harper, F. et al., “Predictors of Answer Quality in Online Q&A Sites,” Proc. of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, pp. 865-874, 2008.

[8] Jeon, J., Corft, W. B., and Lee, J. H., “Finding similar questions in large question and answer archives,” Proc. of the 14th ACM International Conference on Information and knowledge Management, pp. 84-90, 2005.

[9] Zhongbao, K. and Changshui, Z., "Reply networks on a bulletin board system," Physical Review E, Vol. 67, No. 3, 2003.

[10] Adamic, L. A., Zhang, J., Bakshy, E., Ackerman, M. S., "Knowledge sharing and yahoo answers : everyone knows something," Proc. of the 17th International Conference on World Wide Web, pp. 665-674, 2008.

[11] Alessandro, M. et al., "Exploiting Syntactic and Shallow Semantic Kernels for Question/Answer Classification," Proc. of ACL-07, pp. 776-783, 2007.

[12] ODP, http://www.dmoz.org.

[13] Whittaker, S., Terveen, L., Hill, W., and Cherny, L., "The dynamics of mass interaction," Proc. of the 1998 ACM Conference on Computer Supported Cooperative Work, pp. 257-264, 1998.

[14] Search engine guide, http://www.searchengineguide.com/.

[15] Search engine index, http://www.searchenginesindex.com/.

[16] SearchMedica, http://www.searchmedica.com/.

[17] The Porter Stemming Algorithm, http://tartarus.org/~martin/PorterStemmer/.

[18] Twitter, http://twitter.com/.

[19] Vertical search, http://www.verticalsearch.com/.

[20] Virtual Sites, http://www.virtualfreesites.com/.

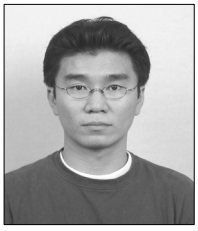
[21] WebMD, http://www.webmd.com/.

[22] Kim, W., Jeong, O.-R., and Lee, S.-W., "On social Web sites," Information Systems, Vol. 35, No. 2, pp. 215-236, 2010.

[23] Yahoo! Answers, http://answers.yahoo.com/.

저 자 소 개

정옥란 (E-mail : orjeong@kyungwon.ac.kr)
2005년 이화여자대학교 컴퓨터공학과 (공학박사)
2005년~2006년 서울대학교 컴퓨터공학부 (박사 후 연구원)
2007년 Univ. of Illinois of Urbana Champaign (visiting scholar)
2008년~2009년 성균관대학교 정보통신공학부 (연구교수)
2009년~현재 경원대학교 소프트웨어 설계·경영학과 (조교수)
관심분야 웹 마이닝, 정보검색, 추천 시스템, 소셜 컴퓨팅



오제환 (E-mail : hide7674@skku.edu)
2006년 성균관대학교 정보통신공학부 대학원 졸업 (석사)
2006년~현재 성균관대학교 정보통신공학부 대학원 박사과정
관심분야 집단지성, 웹마이닝, 추천 시스템, 소셜 컴퓨팅

이은석 (E-mail : eslee@skku.edu)
1985년 성균관대학교 전자공학과 (공학사)
1988년 일본 동북(Tohoku)대학교 대학원 정보공학과 (공학석사)
1992년 일본 동북(Tohoku)대학교 대학원 정보공학과 (공학박사)
1992년~1994년 일본 미쯔비씨 정보전자연구소 특별연구원
1994년~1995년 일본 동북(Tohoku)대학교 Assistant Prof.
1995년~현재 성균관대학교 정보통신공학부 교수
관심분야 소프트웨어공학, 오토노믹/유비쿼터스 컴퓨팅, 에이전트지향 지능형시스템 등