확장된 나이브 베이즈 분류기를 활용한 질문-답변 커뮤니티의 질문 분류

(Modified Naïve Bayes Classifier for Categorizing Questions in Question-Answering Community)

연 종 흠 [†] 심 준 호 [††]
(Jongheum Yeon) (Junho Shim)

이 상 구 [†††]
(Sang-goo Lee)

요 약 소셜 미디어(social media)는 블로그, 소셜 네트워크, 위키 등과 같이 사용자의 참여로 만들어지는 정보 컨텐츠이다. 사용자가 작성한 질문에 다른 사용자들이 답변을 하는 질문-답변 커뮤니티 서비스도 이러한 소셜 미디어의 한 가지로서 지난 몇 년간 많은 양의 정보를 축적해왔다. 하지만 축적된 질문-답변의 양이 많아질수록 이전의 질문을 정확히 검색하는 것은 점점 어려운 작업이 되고 있다. 본 논문에서는 질문-답변 커뮤니티의 효율적인 정보 검색을 위해 확장된 나이브 베이즈 분류기(naïve Bayes classifier)를 이용하여 질문을 그 목적에 따라 정보형, 제안형, 의견형으로 자동 분류하는 기법을 제안한다. 정확한 분류를 위해 분류기는 질문-답변 문서의 구조적인 특징을 활용한다. 실제 질문-답변 커뮤니티의 질문들에 대해 실험을 수행한 결과 71.2%의 분류 정확도를 보였다.

키워드 : 소셜 미디어, 질문-답변 커뮤니티, 나이브 베이즈 분류기, 질문 분류

Abstract Social media refers to the content, which are created by users, such as blogs, social networks, and wikis. Recently, question-answering (QA) communities, in which users share information by questions and answers, are regarded as a kind of social media. Thus, QA communities have become a huge source of information for the past decade. However, it is hard for users to search the exact question-answer that is exactly matched with their needs as the number of question-answers increases in QA communities. This paper proposes an approach for classifying a question into three categories (information, opinion, and suggestion) according to the purpose of the question for more accurate information retrieval. Specifically, our approach is based on modified naïve Bayes classifier which uses structural characteristics of QA documents to improve the classification accuracy. Through our experiments, we achieved about 71.2% in classification accuracy.

Key words : Social Media, Question-Answering Community, Naïve Bayes Classifier, Classifying Questions

1. 서 론

소셜 미디어(social media)는 사용자의 참여로 정보 컨텐츠가 생성되는 온라인 미디어이다. 사용자들은 블로그, 소셜 네트워크, 위키와 같은 소셜 미디어 서비스를 통해 각자의 정보, 의견, 생각 등을 공유한다. 소셜 미디어는 지난 몇 년간 웹 2.0의 관심과 함께 발전하였고, 많은 양의 정보를 축적하게 되었다.

네이버 지식iN, Yahoo! Answers 와 같은 질문-답변 커뮤니티도 이러한 소셜 미디어의 한 가지로서 사용자가 작성한 질문에 대해 다른 사용자가 답변을 하는 게시판 형태의 정보 공유 서비스이다. 질문-답변 커뮤니티는 의견, 조언, 노하우 등과 같이 일반적인 검색엔진으로 쉽게 찾을 수 없는 형태의 정보도 질문을 작성하면 다수로부터 답변을 얻을 수 있다는 장점으로 정보를 얻는 새로운 창구로서 지난 몇 년간 지속적인 주목을 받아왔다. 또한 질문자가 적합한 답변을 선택하고 이에 따라 답변자에게 보상을 주는 시스템은 사용자에게 참여 동기를 부여하면서 답변의 품질을 보장하는 수단으로 작용하고 있다.

검색 포탈들은 이와 같은 질문-답변 커뮤니티에 대한 검색을 '지식검색'이라는 서비스를 통해 제공하고 있다. 하지만 일반적인 웹 문서와 사용자에 의해 생성된 컨텐츠와의 차이점, 질문-답변 컨텐츠의 특수성 등으로 인해

· 본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 육성·지원사업(NIPA-2009-C1090-0902-0031)의 연구결과로 수행되었음
· 이 논문은 2009 한국컴퓨터종합학술대회에서 '확장된 나이브 베이즈 분류기를 활용한 질문-답변 커뮤니티의 질문 분류'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 서울대학교 전기컴퓨터공학부
jonghm@europa.snu.ac.kr
†† 정 회 원 : 숙명여자대학교 정보과학부 교수
jshim@sm.ac.kr
††† 종신회원 : 서울대학교 전기컴퓨터공학부 교수
sglee@europa.snu.ac.kr

논문접수 : 2009년 8월 13일
심사완료 : 2009년 11월 3일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제16권 제1호(2010.1)

높은 신뢰도의 검색 결과 제공은 아직 미흡한 실정이다. 특히, 일반적인 검색 엔진은 웹 문서를 사실 정보를 다루는 문서로서 보고 검색을 수행하기 때문에 질문-답변과 같이 의견, 제안과 같은 사용자의 주관적인 내용이 포함된 문서에 대한 효율적인 검색에는 부적합한 면이 있다. 그러므로 질문-답변 커뮤니티에 대한 검색을 보다 적합하게 수행하기 위해서는 질문-답변을 그 특성에 따라 객관적인 내용을 묻는 질문과 주관적인 의견을 묻는 질문으로 분류해야 하는 필요성이 있다.

본 논문에서는 질문-답변 형식의 문서에서 나타나는 구조적인 특성을 활용하여 질문을 그 목적에 따라 분류하는 방법을 제안한다. 분류는 확장된 나이브 베이즈 분류기(naïve Bayes classifier)를 이용하며, 현재 가장 큰 규모의 질문-답변 커뮤니티 서비스인 Yahoo! Answers의 문서를 이용하여 성능을 측정한다.

논문의 구성은 다음과 같다. 2장에서는 질문-답변 커뮤니티와 관련한 기존 연구를 살펴보고, 3장에서는 질문-답변 서비스의 구조적 특성과 질문의 특성을 설명한다. 4장에서는 문서의 구조적인 특성이 반영된 나이브 베이즈 분류기에 대해 제안한다. 5장에서 분류기의 성능을 측정한 실험 결과에 대해 정리하며, 마지막 6장에서는 결론 및 향후 연구를 기술한다.

2. 관련연구

질문-답변 형태의 문서와 그 특성에 대한 연구는 형태적으로 유사한 유즈넷이나 온라인 게시판을 대상으로 이루어졌다. Whittaker[1]는 유즈넷을 대상으로 사용자의 수, 글의 길이 등의 통계학적인 패턴을 찾아내었으며, Zhongbao[2]는 온라인 게시판의 질문-답변 구조를 이용한 소셜 네트워크 분석을 수행하여 사용자들의 행동 패턴이 그들의 관심 공간에 따라 달라지는 것을 보였다.

2000년대 들어서 서비스가 상용화된 이후 질문-답변 커뮤니티에 특성화된 연구가 본격화되었는데, 대표적으로 질문-답변 문서를 그 대상으로 하여 정보 검색의 성능을 높이거나 문서 품질 측정 방법을 제시한 연구들이 있다. Jeon[3]은 답변의 유사도를 통해 의미적으로 비슷한 질문을 구한 후, 이들에서 관련 있는 키워드들을 추출하여 검색의 성능을 향상 시켰다. Agichtein[4]은 질문-답변 문서의 길이, 구두점의 수와 같은 텍스트 정보 뿐만 아니라 답변자의 등급, 추천수, 조회수 등의 비텍스트 정보들을 사용하여 개선된 문서의 품질 측정 방법을 제안하였다. 국내의 연구로 Lee[5]는 네이버 지식iN을 대상으로 문서 내용의 신뢰도를 측정하기 위해 텍스트 정보 기반의 문서 신뢰도 자질을 정의하였으며, 이를 사용한 문서 품질 평가 모델을 제안하였다.

한편 질문을 특정 기준에 따라 분류하여 그 특성을 분석한 연구들이 있다. Park[6]은 답변의 신뢰도를 평가하기 위해 질문을 객관적인 근거를 포함하는 답변을 요구하는 지식형 질문과 속담이나 생활지식 등 학문적 근거는 없지만 상식적인 것을 묻는 생활형 질문으로 나누었다. Adamic[7]은 Yahoo! Answers의 카테고리들을 K-Means 알고리즘을 사용하여 문서의 길이, 질문당 답변의 개수, 사용자간의 상호작용 패턴 등의 특성에 따라 세 가지 분류로 클러스터링 하였다. 각각의 클러스터가 포함하고 있는 카테고리에 속하는 질문들의 특성을 분석한 결과 토론형 질문, 상식, 조언형 질문, 사실형 질문 중 어느 질문을 많이 포함하고 있는지에 따라 다른 클러스터에 포함되는 것으로 나타났다. Kim[8]은 사용자가 어떤 기준으로 가장 좋은 답변을 선택하는지에 대한 통계적인 분포를 구하였다. 이때 질문을 정보형(Information), 제안형(Suggestion), 의견형(Opinion) 질문과 어느 분류에도 속하지 않는 기타(Others) 질문으로 나누었다.

이와 같이 질문-답변 커뮤니티와 관련하여 많은 연구가 활발히 진행되고 있다. 하지만 질문-답변 문서 구조나 질문의 목적과 같은 특성을 반영한 정보 검색에 대한 연구는 미비한 실정이다. 앞서 살펴본 연구들도 질문-답변을 기존의 정보 검색의 대상이 되는 문서와 동일시하거나, 질문-답변 커뮤니티에서 수집되는 비텍스트적 정보를 활용하는 정도에 그치고 있다.

3. 질문-답변 커뮤니티 특성

질문-답변 커뮤니티는 사용자가 작성한 질문에 대해 다른 사용자들이 답변을 한 후, 질문자가 답변들 중 가장 적절한 답변 하나를 선택한다. 질문자로부터 선정된 답변을 작성한 사용자는 보상 개념의 점수를 얻게 된다. 그림 1은 질문-답변 서비스 중 가장 큰 규모인 Yahoo! Answers의 질문과 답변의 한 예시이다.

Yahoo! Answers는 상위 26개의 카테고리와 약 1000여 개의 하위 카테고리로 이루어져 있다. 질문은 'Beauty & Style' 카테고리와 같이 일상적인 질문 뿐만 아니라 'Science & Mathematics' 카테고리와 같이 비교적 전문적인 내용의 질문들도 포함한다.

3.1 질문 분류

질문을 그 목적에 따라 분류하고자 할 때 카테고리의 주제를 기준으로 나누는 것은 적절하지 않다. 질문의 목적은 카테고리의 주제 따라 특정 목적이 다수를 차지하기도 하지만 카테고리에 반드시 의존적이지는 않음을 보였다. Yahoo! Answers의 가장 많은 수의 질문이 있는 카테고리들(2009년 4월 기준)은 표 1에서와 같이 'Entertainment & Music', 'Family & Relationships', 'Society & Culture'순으로 나타났다. 이 카테고리들의

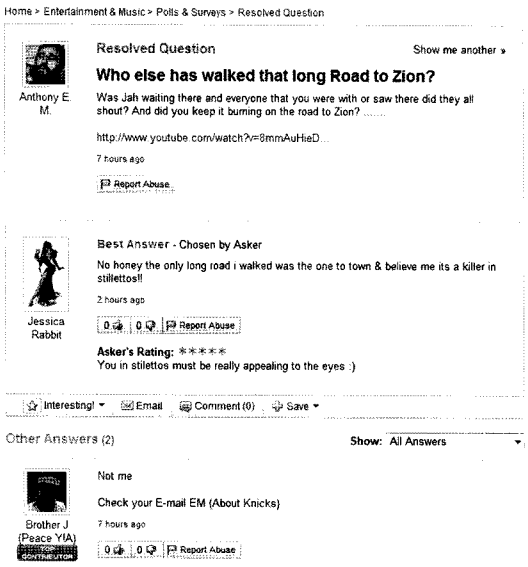

그림 1 Yahoo! Answers의 질문과 답변

표 1 Yahoo! Answers의 질문 수 상위 10개 카테고리

카테고리	질문	백분율
Entertainment & Music	9131480	15.7
Family & Relationships	5585199	9.6
Society & Culture	4335511	7.4
Health	4232218	7.3
Computers & Internet	3502627	6.0
Science & Mathematics	2772160	4.8
Politics & Government	2654574	4.6
Beauty & Style	2606112	4.5
Sports	2545362	4.4
Education & Reference	2435767	4.2

경우 "Do you want to play outside in the rain?"와 같이 의견을 묻는 질문이 비교적 많은 수를 차지하고 있었지만, "Figure out these 2 songs name please?"와 같이 객관적인 정보를 요구하는 질문도 존재하였다. 마찬가지로 'Computer & Internet' 카테고리의 경우 "How do I delete all the cookies on my laptop?"와 같이 정보를 구하는 질문과 함께 "What graphics card is better?"와 같이 의견을 구하는 질문도 많은 비율을 차지하였다.

한편, 질문을 그 목적에 따라 분류하는 기준은 앞서 살펴본 관련 연구[6-8]와 같이 여러 가지가 있다. 본 연구에서는 다른 연구들에 비해 상대적으로 명확한 기준을 제시한 [8]의 정보형, 제안형, 의견형의 분류를 사용하였으며 기타는 제외하였다. 정보형은 구체적인 사실이나 현상의 이해를 묻는 질문으로 "What is the voltage and capacitance?"와 같은 예가 있다. 제안형은 조언, 추천, 실용적 해결책을 구하는 질문으로 "I have 3 days to visit NYC, what places I shouldn't miss?"와 같은 질문이 속한다. 의견형은 사회적 이슈에 대해 다른 사람들의 생각이나 취향을 묻거나 토론을 요구하는 질문들로 "What are five simple things that's America?"와 같은 것들이다.

3.2 질문-답변 문서의 구조 및 활용

질문-답변 문서는 텍스트로 이루어진 문서이지만, 그림 1과 같이 '질문', '선택된 답변', '그 외 답변들'의 세 가지 부분으로 구성된다. 또한 질문은 제목과 본문으로 구성되며 답변들은 본문만으로 이루어져 있다. 이를 이용하여 질문-답변 문서를 속성-값의 형태로 표현할 수 있다. 속성에 해당하는 것은 '질문 제목', '질문 본문', '선택 답변 본문', '그 외 답변 본문'이고, 값에 해당하는 것은 각각의 텍스트를 구성하는 키워드들의 집합이 된다. 즉 질문-답변 문서는 일종의 준구조적 문서(semi-structured document)로 볼 수 있다.

질문의 목적에 따른 문서의 분류 시, 이와 같은 속성-값의 형태로 질문-답변 문서가 표현된다는 점을 이용할 수 있다. 나이브 베이즈 분류기는 불연속 값을 갖는 속성-값 쌍의 집합을 분류하기 위해 고안된 확률이론 기반의 분류 알고리즘이다[9]. 일반적으로 나이브 베이즈 분류기는 이메일이나 전자 카탈로그와 같은 텍스트 문서의 분류에 널리 쓰이고 있다. 질문-답변 문서도 텍스트 문서이고, 속성-값의 쌍으로 표현이 가능하기 때문에 분류 방법으로 나이브 베이즈 분류기를 적용할 수 있다. 또한 각각의 속성은 같은 값이 존재 한다고 하더라도 그 분포나 중요도에 따라 문서의 분류에 미치는 영향이 달라지기 때문에, 이를 고려한 확장된 나이브 베이즈 분류기[10]를 이용한다.

일반적으로 의견과 같이 주관적인 표현이 포함된 문서를 분류하는데는 다른 문서 분류 알고리즘보다 SVM과 나이브 베이즈 분류기의 정확도가 높다고 알려져 있다[11]. 하지만, SVM의 경우 분류 대상이 두 가지 보다 많을 경우 알고리즘을 직접 적용하기 어려운 점이 있다. 따라서 질문-답변 문서의 분류에는 나이브 베이즈 분류기를 적용하는 것이 적절하며, 실험에서 일반적인 나이브 베이즈 분류기와 확장된 나이브 베이즈 분류기의 정확도를 비교하였다.

4. 문서 분류

4.1 확장된 나이브 베이즈 분류기

텍스트를 분류하는 나이브 베이즈 분류기는 분류들의 집합 C, 텍스트의 각 단어의 위치인 속성의 집합 $<a_1, a_2, ..., a_n>$, 텍스트를 구성하는 단어들의 리스트의 집합을 값으로 하는 $<v_1, v_2, ..., v_n>$이 주어졌을 때 다음 식

에 의해 가장 높은 사후 확률(posterior probability)을 갖는 분류를 선택한다.

$$C_{NB} = \underset{c_j \in C}{argmax}\, P(c_j) \prod_i P(a_i = v_i | c_j)$$

이때, 각 속성은 같은 도메인의 단어들의 집합으로 이루어지고 그 분포가 같다고 가정한다. 이러한 가정은 평이한 텍스트(plain text)에는 합당하나, 구조화 된 문서에 반드시 적용되지는 않는다. 가령 'recommendation'이라는 단어가 질문 제목과 본문에서 나올 때의 조건부 확률 $P(a_i = v_i|c_j)$은 다를 것이다.

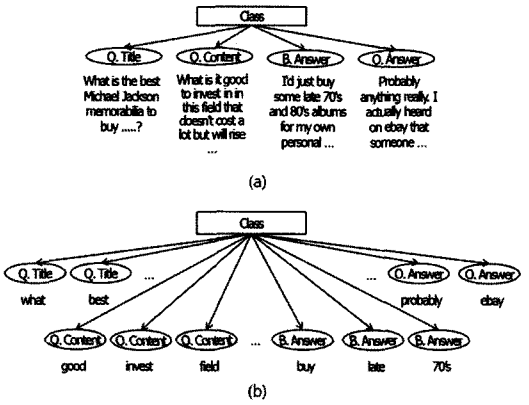

그림 2 확장된 속성 예시

그림 2는 3.2절에서 설명한 질문-답변을 속성-값의 형태로 표현한 모델(a)과 그것을 키워드를 기준으로 확장한 모델(b)의 예시이다. 각 속성의 값은 텍스트로 이루어진다. (a)는 속성의 값으로 전체 텍스트를 사용하므로 텍스트와 정확히 일치하는 값만이 분류의 결과에 영향을 준다. 하지만 학습된 데이터와 입력된 문서 사이에 일부의 단어만 일치할 때의 확률도 고려를 해야 하므로 (b)와 같이 키워드 형태로 확장된 모델이 필요하다. 또한 질문-답변 문서는 구어체의 문장이 다수 포함된 문서로 키워드 집합을 구성할 때 파서를 활용한 키워드 추출 및 스테밍(stemming) 작업이 필요하다.

추출된 키워드를 t_{ik}라 할 때 속성의 값과 확률 계산 식은 다음과 같이 표현된다.

$$v_i = \{t_{i1}, t_{i2}, \ldots, t_{im}\}$$

$$P(a_i = v_i|c_j) = \prod_k P(t_{ik}\ appears\ in\ a_i|c_j)$$

$$= \prod_k \frac{n(c_j, a_i, t_{ik})}{n(c_j, a_i)}$$

이때, $n(c_j, a_i, t_{ik})$는 문서 속성 a_i가 분류 c_j에 속할 때의 키워드 t_{ik}의 빈도수이고, $n(c_j, a_i)$는 분류 c_j에 속하는 모든 문서의 속성 a_i가 포함하는 모든 키워드 빈도수의 합이다. 또한 분류 시 입력된 키워드가 학습 데이터에 존재하지 않을 때의 최소값은 $1/n(c_j, a_i)$을 사용한다. 이에 따라 최종적으로 얻어진 식은 다음과 같다.

$$C_{NB} = \underset{c_j \in C}{argmax}\, P(c_j) \prod_{i,k} P(t_{ik}\ appears\ in\ a_i|c_j)$$

$$= \underset{c_j \in C}{argmax} |c_j| \left\{ \prod_{i,k} \frac{n(c_j, a_i, t_{ik})}{n(c_j, a_i)} \right\}$$

$|c_j|$는 분류에 속하는 문서의 수이다.

4.2 속성의 정규화 및 가중치 부여

확장된 속성을 이용한 분류는 속성의 길이에 따라 분류에 미치는 영향이 달라지는 문제점이 있다. 예를 들어 10단어로 구성된 질문과 30단어로 구성된 답변의 문서가 있을 때, 속성의 값으로 질문보다 답변이 더 많은 키워드를 포함하게 된다. 일반적으로 질문이 답변보다 분류에 더 많은 영향을 미치지만, 이와 같은 상황에서는 더 많은 키워드로 이루어진 답변의 영향력이 더 크게 된다. 따라서 속성의 정규화를 통해 각 속성들이 미치는 영향력을 동일하게 한 후, 중요도에 따라 가중치를 부여해야 한다. 정규화와 가중치가 적용된 식은 다음과 같다.

$$C_{NB} = \underset{c_j \in C}{argmax} |c_j| \left\{ \prod_{i,k} \frac{n(c_j, a_i, t_{ik})}{n(c_j, a_i)} \right\}^{\frac{w_i}{|v_i|}}$$

여기서 $|v_i|$는 속성값 v_i의 키워드의 개수이고, w_i는 가중치를 나타낸다. w_i는 실험적으로 가장 적절한 값을 찾는다.

5. 실험 및 분석

실험에 사용한 데이터는 Yahoo! Answers에서 답변 작성이 완료된 질문들을 대상으로 수집하였다. 데이터는 'Family & Relationships', 'Computers & Internet', 'Sports' 카테고리에서 2009년 3월 20일부터 22일 까지 작성된 669개의 질문과 2176개의 답변을 포함하였다. 이를 수작업으로 질문 목적에 따라 분류 한 후 10-집단 교차검증(10-fold cross validation)을 이용하여 검증하였다.

실험은 임의 분류, 일반 나이브 베이즈 분류기, 확장된 속성 기반 분류, 정규화 반영 분류, 정규화 및 가중치 반영 분류에 대해 실시하였다.

표 2 질문 목적에 따른 데이터 분류

질문 목적	개수	비율(%)
정보형(Information)	113	16.9
제안형(Suggestion)	282	42.1
의견형(Opinion)	274	41.0
계	669	100

표 3 실험에 사용된 분류 기법

Random	임의 분류
Flat	일반 나이브 베이즈 분류
Modified	확장된 속성 기반 분류
Normalized	정규화 반영 분류
Weighted	가중치 반영 분류

실험 결과는 그림 3과 같다. 일반 나이브 베이즈 분류기의 경우 약 64%의 정확도를 보였으며, 확장된 속성 기반 분류는 이와 비슷한 결과를 보였다. 여기에 정규화 과정과 가중치 정보를 추가적으로 사용할 경우 분류 정확도는 각각 69%, 71%로 더 높아졌다. 이 실험에서 속성별 가중치는 '질문 제목', '질문 본문'에 1.2, '선택 답변 본문', '그 외 답변 본문' 에 1의 비율로 할당하였다. 하지만 가중치를 이상의 비율로 변화시켜도 상대적으로 분류 정확도가 민감하게 반응하지는 않았다.

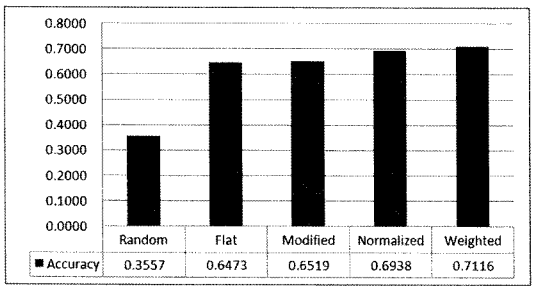

그림 3 정확도 측정 결과

6. 결론 및 향후 연구

본 논문은 질문-답변 문서의 구조적인 특징을 이용하여 문서를 질문의 목적에 따라 자동으로 분류하는 방법에 대해 다루었다. 이를 위해 질문-답변 커뮤니티의 특징을 설명하고, 확장된 나이브 베이즈 분류기에 문서의 구조적인 특징을 반영하는 방법에 대해 설명하였다.

향후 질문-답변 커뮤니티에 대한 정보 검색에 있어 분류된 문서에 따라 다른 검색 방법을 적용하여 검색의 성능을 높이는 연구를 지속할 예정이다. 또한, 문서 분류시 구조적 특징과 같은 텍스트 정보뿐만 아니라 사용자의 평판 등과 같은 비텍스트 정보를 활용하여 정확도를 높이는 것도 가치 있는 연구가 될 것이다.

참 고 문 헌

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

[2] K. Zhongbao, Z. Changshui, "Reply networks on a bulletin board system," *Physical Review E*, http://pre.aps.org/abstract/PRE/v67/i3/e036117

[3] J. Jeon, W.B. Croft, J.H. Lee, "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.

[4] E. Agichtein, C. Castillo, D. Donato, A. Gionis, G. Mishne, "Finding high-quality content in social media," *Proc. of the International Conference on Web Search and Web Data Mining*, pp.183-194, 2008.

[5] J. Lee, Y. Song, H. Rim, "Quality Prediction of Knowledge Search Documents Using Text-Confidence Features," *Proc. of the 19th Annual Conference on Human and Cognitive Language Technology*, pp.62-67, 2007. (in Korean)

[6] S. Park, J. Lee, J. Jeon, "Evaluation of the documents from the Web-based Question and Answer Service," *Journal of the Korean Society for Library and Information Science*, vol.40, no.2, pp.299-314, 2006. (in Korean)

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

[8] S. Kim, J. S. Oh, S. Oh, "Best-answer selection criteria in a social Q&A site from the user-oriented relevance perspective," *Proc. of the American Society for Information Science and Technology*, vol.44, no.1, pp.1-15, 2007.

[9] T. Mitchell, Machine Learning, McGraw-Hill, 1997.

[10] Y. Kim, T. Lee, J. Chun, S. Lee, "Modified Naïve Bayes Classifier for E-Catalog Classification," *Lecture Notes in Computer Science*, vol.4055, pp.246-257, 2006.

[11] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," *Proc. of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, pp.79-86, 2002.