커뮤니티 기반 Q&A서비스에서의 질의 할당을 위한 이용자의 관심 토픽 분석에 관한 연구

A Study on Mapping Users' Topic Interest for Question Routing for Community-based Q&A Service

박종도 (Jong Do Park)*

초 록

본 연구에서는 커뮤니티 기반 질의응답 서비스에서의 질의할당을 위하여, 해당 커뮤니티에 축적된 질의응답 데이터 세트를 이용하여 해당 카테고리내의 토픽을 분석하고 이를 바탕으로 해당 토픽에 관심을 가지는 이용자의 관심 토픽을 분석하고자 하였다. 특정 카테고리 내의 토픽을 분석하기 위해서 LDA기법을 사용하였고 이를 이용하여 이용자의 관심 토픽을 모델링하였다. 나아가, 커뮤니티에 새롭게 유입되는 질의에 대한 토픽을 분석한 후, 이를 바탕으로 해당 토픽에 대해 관심을 가지고 있는 이용자를 추천하기 위한 일련의 방법들을 실험하였다.

ABSTRACT

The main goal of this study is to investigate how to route a question to some relevant users who have interest in the topic of the question based on users' topic interest. In order to assess users' topic interest, archived question-answer pairs in the community were used to identify latent topics in the chosen categories using LDA. Then, these topic models were used to identify users' topic interest. Furthermore, the topics of newly submitted questions were analyzed using the topic models in order to recommend relevant answerers to the question. This study introduces the process of topic modeling to investigate relevant users based on their topic interest.

키워드: 커뮤니티 기반 Q&A, 소셜 Q&A, 지식검색 커뮤니티, 질의할당, 토픽 모델, LDA (Latent Dirichlet Allocation)
community-based Q&A, social Q&A, question routing, question triage, topic model, LDA (Latent Dirichlet Allocation)

---

* 중앙대학교 문헌정보학과 강사(dlibrary@gmail.com)

■ 논문접수일자 : 2015년 8월 31일 ■ 최초심사일자 : 2015년 9월 11일 ■ 게재확정일자 : 2015년 9월 14일
■ 정보관리학회지, 32(3), 397-412, 2015. 〔http://dx.doi.org/10.3743/KOSIM.2015.32.3.397〕

1. 서 론

1.1 연구의 필요성 및 의의

어린아이에서부터 노인에 이르기까지 인간에게 있어서 자신 또는 주변의 문제들을 해결하기 위해 주변 사람들에게 묻고 답하는 일련의 행태는 매우 자연스러운 행동이다. 전통적으로 도서관은 도서관이 축적해 온 지식을 바탕으로 참고사서나 문헌을 통해 이용자들의 궁금증에 대한 해답 또는 해답을 얻을 수 있는 방법을 제공하기 위해 애써왔다. 그러나 지난 10년동안 인터넷과 같은 정보기술의 발달과 소셜 네트워크 서비스의 폭발적인 인기로 인해 많은 이용자들이 도서관을 떠나 웹으로 향했고 사이버 공간에서 커뮤니티를 형성하여 다른 이용자들과 함께 소통하기 시작하였다. 특히 네이버 지식인, Yahoo Answers, AskMetafilter 등과 같은 커뮤니티 기반의 질의응답 서비스는 이용자들이 사이버 공간에서 커뮤니티를 형성하고 그 안에서 질의응답의 형식으로 서로 소통할 수 있는 기반을 제공하여 왔다. 이와 같은 서비스의 편의성으로 인해 점점 더 많은 이용자들이 커뮤니티 기반 질의응답 서비스를 이용하게 되었고 결과적으로 매일매일 처리해야 하는 질의의 양도 폭발적으로 증가하였다. 그 결과, 이용자의 질의 수가 급증하는 것과 함께 커뮤니티 내에서 다른 이용자로부터 일정한 기일 내에 답변을 제공받지 못하는 무응답 질의의 수도 급증하여 전반적으로 해당 서비스의 질에 부정적인 영향을 미치게 되었다. 이러한 문제를 해결하기 위하여 많은 연구자들이 Yahoo Answers와 같은 서비스를 중심으로 다양한 연구들을 수행하였는데, 이러한 연구는 크게 시스템을 중심으로 한 정보검색적인 연구주제와 이용자를 대상으로 한 이용자의 정보추구행태에 대한 연구로 구분할 수 있다. 정보검색적인 연구주제는 주로 시스템을 활용하여 주어진 질의에 대해 적합하다고 평가되는 이용자 혹은 가장 좋은 답변을 제공할 수 있는 전문가(best answerer)를 검색하기 위한 다양한 방법들을 시도하였다. 한편, 커뮤니티기반 질의응답 서비스를 이용한 기존의 이용자 중심의 연구들은 소셜 네트워크 서비스에 참여하는 이용자의 참여 동기, 정보활용 등에 대해 연구하였다. 최근에는 OCLC와 몇몇 연구자를 중심으로 기존의 도서관 참고서비스에 커뮤니티 기반 질의응답 서비스를 연계하는 새로운 형태의 참고서비스를 개발하기 위한 연구가 수행되기도 하였다.

본 연구에서는 커뮤니티 기반 질의응답 서비스의 품질을 향상시키기 위한 방법의 하나로 커뮤니티에 포스팅된 질의에 대해 적합한 이용자를 시스템이 추천하는 질의할당의 방법을 제시하고자 한다. 본 연구에서는 커뮤니티 기반 질의응답 서비스에서의 질의할당을 위해 이용자의 관심 토픽을 분석하여 이를 활용하는 방법을 제시하고자 한다.

2. 선행 연구

이용자의 참여를 주된 특징으로 하는 웹 2.0의 등장으로 인해 촉발된 소셜 네트워크 서비스는 해당하는 커뮤니티의 구성원인 이용자의 참여와 기여도에 크게 의존하고 있다. 따라서, 연구자들이 소셜 네트워크 서비스의 주요 동력

원인 이용자에 대해 관심을 가지고 연구를 진행하였다. 다른 한편, 일반적인 웹 기반 정보 서비스는 그 서비스의 제공 결과로 생성된 정보를 쉽게 수집, 축적할 수 있으므로 몇몇 연구자들은 커뮤니티 기반 질의응답 서비스의 결과로 수집, 축적된 정보를 활용하여 연구하였다.

2.1 이용자의 정보추구 행태에 관한 연구

많은 연구자들이 커뮤니티 기반 질의응답 사이트를 대상으로 이용자의 정보추구 행태에 관한 연구를 수행하였다. Adamic, Zhang, Bakshy, Ackerman(2008)은 Yahoo Answers를 대상으로 한 연구에서 이용자들이 Yahoo Answers의 여러 카테고리에 걸쳐서 자신들의 지식을 공유하는 데 있어서 각각 다른 차이를 보이는 것을 관찰하였다. Shah(2011)는 Yahoo Answers에서 답변을 받는데 걸리는 응답 시간이 해당 서비스의 효율성과 이용자 만족도에 미치는 영향을 연구하여 더 빨리 응답한 답변이 질의자에 의해 가장 만족할 만한 답변(best answer)으로 채택될 가능성이 높음을 발견하였다.

2.2 이용자의 동기에 대한 연구

최근 수 년동안 소셜 네트워크 서비스가 급성장하였는데, 몇몇 연구자들은 소셜 네트워크 서비스를 이용하여 이용자들이 자신들의 지식을 다른 이용자들과 서로 공유하고 또 서로 협업하는 동기에 관심을 가졌다. Farzan과 Brusilovsky(2010)는 소셜 네트워크 서비스를 이용하는 이용자의 동기를 크게 내부적인 동기(intrinsic motivation), 외부적인 동기(extrinsic motivation)로 구분하고 내부적인 동기는 외부나 다른 이용자들로부터의 보상 없이 자발적으로 활동에 참여하는 것으로 설명하고 외부적인 동기는 다른 외부로부터의 보상을 기대하고 활동에 참여하는 것으로 설명하였다. Oh(2010)는 커뮤니티 기반 질의응답 서비스에 참여하는 이용자의 동기를 즐거움(enjoyment), 자기 효능감(self efficacy), 학습, 개인적인 이득 등의 개인적인 요인(personal factors)과 이타심, 커뮤니티에 대한 관심, 공감(emphaty), 명성, 일반적인 호혜성(generalized reciprocity) 등의 사회적인 요인(social factors)로 구분하여 설명하였다.

2.3 이용자 식별에 관한 연구

몇몇 연구자들이 커뮤니티기반 질의응답 서비스를 대상으로한 이용자들에 관심을 갖고 연구한 한편, 또다른 부류의 연구자들은 해당 사이트들을 대상으로 주어진 질의에 대한 답변자를 식별하기 위한 연구들을 수행하였다. 이용자를 식별하기 위한 이러한 부류의 연구에서 연구자들은 이용자 식별과 관련하여 서로 용어들을 사용하고 있는데, 그 예는 베스트 답변자(best answerer) (Bouguessa, Dumoulin, & Wang, 2008; M. Liu, Liu, & Yang, 2010; Qu, Qiu, He, & Zhang, 2009), 전문가(expert) (Zhang, Tang, & Li, 2007), 권위있는 이용자(authoritative user) (Bouguessa, Dumoulin, & Wang, 2008; Agichtein, Castillo, Donato, Gionis, & Mishne, 2008; Jurczyk & Agichtein, 2007) 등의 용어들이다. Zhang, Tang, Li(2007)는 이용자의 질문 수와 답변 수를 활용하여 이를 하나의 수치로 표현한 Z-score를 제안하여 이용자의 전

문성을 상대적으로 평가하여 이를 이용자들 가운데 상대적인 전문성을 가진 이용자를 식별하고자 하였다. Bouguessa, Dumoulin, Wang(2008)은 권위있는 이용자(authoritative user)를 식별하기 위하여 링크 분석을 시도하였으며 각 이용자가 제공한 베스트 답변의 숫자가 권위있는 이용자의 식별에 영향을 미치는 주요한 요인임을 발견하였다.

2.4 유사 질의 식별에 관한 연구

몇몇 연구자들은 커뮤니티 기반 질의응답 사이트를 대상으로 유사한 질의를 식별하는 연구를 진행하였다. 이러한 연구는 기존의 서비스를 통해 축적된 질의응답 아카이브에서 유사한 질의의 존재 여부를 탐색하게 함으로써 이용자들이 아카이브로부터 기존에 제공된 답변을 먼저 살펴보게 하여 중복된 질문이 해당 커뮤니티에 포스팅되는 것을 예방하는 효과를 거둘 수 있다. Jeon, Croft, Lee(2005)는 유사한 질의를 탐색하기 위한 방법으로 번역에 기반한 검색모델을 제시하였다. 최초의 원질문을 다른 언어로 번역하고 이를 다시 원질문의 언어로 번역하는 동안 본래 원질문에 포함된 단어들이 번역 모델의 확률에 의해 키워드의 확장이 일어난다. 이를 바탕으로 기존의 아카이브로부터 원 질문과 유사한 질문을 검색하고자 시도하였다.

2.5 질의 할당에 관한 연구

Chang과 Pal(2013)은 커뮤니티기반 질의응답 서비스를 위한 질의할당에 관한 연구를 수행하였다. 이들은 시스템에 유입되는 새로운 질의를 대체가능한 이용자 그룹(팀)에 할당하기 위한 추천시스템을 제안하고 주어진 질의에 대해 대체가능한 이용자 그룹을 추천하기 위한 평가지표로 호환가능성(compatibility), 이용가능성(availability), 이용자의 전문성 등을 고려하였다. 이용자의 전문성을 평가하기 위하여 이용자의 토픽 전문성을 모델링하였는데 이때 질문의 토픽을 모델링하기 위하여 Spectral Cluster(SC) 기법(Ng, Jordan, & Weiss, 2002)과 Latent Dirichlet Allocation(LDA) 기법을 활용하여 토픽 모델링의 성능을 비교하여 SC기법이 질문의 토픽을 모델링하는데 있어서 더 나은 성능을 나타내는 것을 관찰하였다. 이는 해당 연구자들이 질의의 토픽을 모델링하기 위하여 이용자에 의해 기입된 태그 정보를 추가로 활용하였기 때문이다.

2.6 신뢰성에 대한 연구

커뮤니티 기반 질의응답 사이트의 인기와 함께 증가하는 질의의 수와 이용자의 증가로 인해 해당 커뮤니티 내에서 제공되는 정보에 대한 신뢰성이 중요한 연구 주제가 되었다. Kim(2012)은 Yahoo Answers나 네이버 지식인과 같은 지식검색 커뮤니티와 관련된 신뢰성 문제에 대한 선행연구들을 국내외의 문헌을 조사하여 정리하고 향후 연구과제를 제시하였다.

3. 연구설계 및 방법

3.1 데이터 수집

본 연구의 실험을 수행하기 위해서 실제 서비스되고 있는 커뮤니티 기반 질의응답 사이트 중

의 하나인 Ask Metafilter로부터 수집된 데이터를 사용하였다. 기존의 많은 연구자들이 주로 Yahoo Answers의 데이터를 주로 사용하였는데 반해, 본 연구에서 위의 상대적으로 적은 규모를 지닌 Ask Metafilter로부터 수집한 데이터를 연구에 활용하였다. 그 주된 이유는 Ask Metafilter 사이트가 Yahoo Answers에 비해 질의응답의 수나 이용자의 규모에 있어서 상대적으로 작은 규모이지만, 그럼에도 불구하고 해당 사이트의 서비스가 커뮤니티 기반 질의응답 서비스들이 출범하기 시작한 초창기인 2003년에 시작된 점, 데이터를 완전하게 수집할 수 있는 점, 이용자의 프로파일 정보에 대한 완전한 접근 등의 장점이 있어 연구대상으로 선정하였다. 수집한 데이터의 수록 기간은 2003년 12월에서부터 2010년 11월까지 해당 사이트에 축적된 모든 질의응답 자료들을 수집하였다. 전체 데이터 세트의 요약 정보는 〈표 1〉과 같다.

〈표 1〉 Ask Metafilter 2011년도 데이터 세트 요약

이용자 수	23,375
질의자 수	14,448
응답자 수	18,514
질의 수	95,139
응답 수	1,129,284
카테고리 수	20
수록기간	2003.12.1~2010.11.30

본 실험연구에서는 위의 데이터 세트중에서 질의 수가 가장 많은 상위 2개 카테고리인 'Computers and Internet(CI)'와 'Media Arts (MA)'를 대상으로 선정하여 연구를 진행하였다. 실험에 사용된 데이터 세트의 요약 정보는 〈표 2〉와 같다.

〈표 2〉 실험 데이터 세트 요약 정보

	CI	MA
질의자 수	6,570	4,826
응답자 수	10,168	11,091
질의 수	17,969	9,161
응답 수	128,280	114,232

3.2 데이터세트 준비

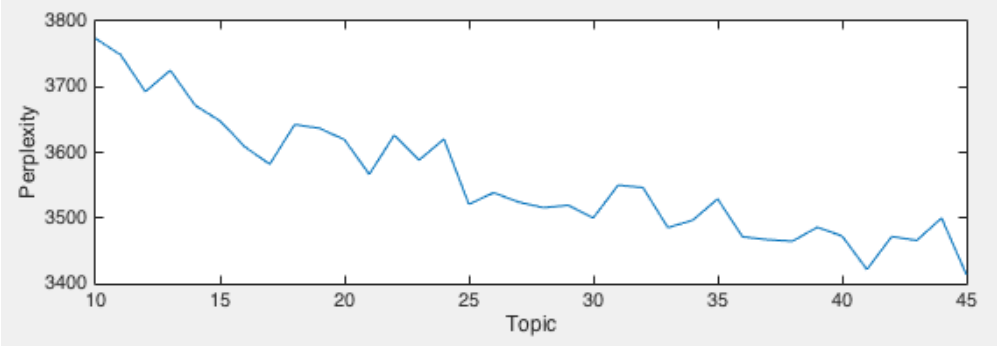
본 연구에서 사용한 데이터세트는 Ask Metafilter 사이트로부터 수집한 질의응답의 세트이다. 우선 이 연구의 주요 과제인 질의응답 세트의 토픽을 모델링하고 이의 품질을 평가하기 위하여 실험 데이터세트를 학습용 데이터세트(training dataset)과 평가용 데이트세트(testing dataset)의 두 그룹으로 7:3의 비율로 분할하였다. 학습용 데이터세트에는 질의와 이에 대한 응답의 내용이 모두 포함되었으나 평가용 세트에는 질의의 내용만 포함하였다. 학습용 데이터세트와 평가용 데이터세트의 구성을 '질의+응답(학습용 데이터세트)'과 '질의(평가용 데이터세트)'로 다르게 한 이유는 커뮤니티기반 질의응답 사이트를 통해 실제로 서비스된 데이터를 활용하여 기계학습을 하는 시점에는 해당 커뮤니티의 질의응답 시스템에 포스팅된 질의응답의 내용을 모두 활용할 수 있지만, 기계학습이 완료된 후, 새로 포스팅되는 질문을 처리하는 미래의 시점에는 해당 질문에 대한 다른 이용자의 응답은 블랙박스처럼 알 수 없기 때문이다. 따라서, 본 연구에서는 우선 학습용 데이터세트를 이용해 기계학습을 하여 질의응답에 대한 토픽 모델을 생성하고 이 모델을 바탕으로 새로 유입되는 질의에 대해 해당 질의의 토픽을 추정하고자 하였다.

3.3 데이터 처리

주어진 데이터세트를 이용해 토픽을 모델링을 하기 위해서 시스템은 수많은 계산을 한다. 본 연구에서는 시스템의 처리 속도를 향상시키기 위하여 두 글자 이하의 영어 단어를 불용어로 간주하여 처리 대상에서 제외하였고, 아울러 사전에 정의된 불용어 사전을 적용하여 불용어 사전에 포함된 단어들도 제외하였다. 또한 전체 데이터세트에서 출현빈도수가 높은 30개의 단어들도 불용어에 포함하였다.

3.4 토픽 모델링

커뮤니티 사이트에 새롭게 유입된 이용자의 질의를 해당 질의에 관심이 있을 것으로 예상되는 이용자에게 할당하기 위한 한 방법으로 본 연구에서는 LDA 기법을 응용하였다. 본 연구에서 주어진 질의에 대해 응답할 것으로 예상되는 이용자를 식별해 내기 위해 수집한 데이터세트에서 질의와 응답의 짝으로 이루어진 데이터를 하나의 문서로 간주하여 해당 문서의 토픽을 분석하고, 이를 바탕으로 해당 문서 내의 질의와 답변을 직접 수행한 이용자의 관심 토픽을 추정하고자 하였다. LDA를 기반한 토픽 모델링을 하기 위하여 Stanford Topic Modeling Toolbox를 사용하였다. LDA기법의 성능은 사전에 알고 있는 토픽의 갯수의 결정에 크게 영향을 받는데, 실제 데이터에서는 각 카테고리에 속해있는 하위 토픽의 갯수를 알지 못하므로 이를 결정하는 것이 매우 중요하다. 주어진 카테고리의 하위 토픽의 갯수를 결정하기 위해 본 연구에서는 각 카테고리별로 최소 10개에서 최대 45개에 이르는 토픽을 사전에 결정하고 이 중 perplexity의 값을 최소화하는 토픽의 수를 채택하였다. 물론 토픽의 갯수를 더 많이 증가시킬수록 perplexity의 값이 점점 더 줄어드는 것이 자연스럽게 관찰되지만, 실제로 커뮤니티 기반 질의응답 사이트에서 하나의 카테고리 내에서 45개 이상의 하위 토픽 또는 서브 카테고리가 관찰되는 경우가 낮으므로 선정된 범위 내에서 가장 최상의 토픽 모델을 제시하는 토픽의 수를 결정하였다. 이러한 결정과정을 거쳐 CI 카테고리는 44개, MA 카테고리는 45개의 하위 토픽으로 토픽의 갯수를 정하고 모델링하였다.



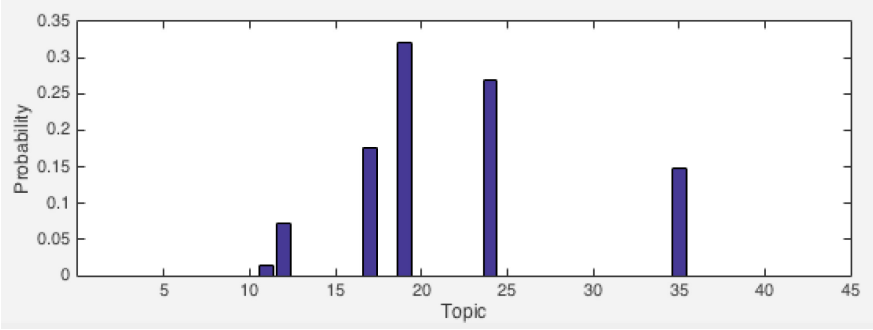
〈그림 1〉 토픽 수와 Perplexity의 상관 관계 - MA 카테고리의 예

4. 연구결과 및 분석

4.1 토픽 추출

본 실험 연구를 위해 먼저 사전에 구분해 놓은 학습용 데이터세트를 Stanford Topic Modeling Toolbox[1]를 활용하여 학습용 데이터세트에 포함된 모든 질의응답 세트에 대해 툴박스 내의 LDA 기능을 활용하여 토픽 추출을 시도하였다. 위 툴박스의 LDA 모듈은 주어진 데이터세트를 사전에 정의한 n개의 토픽으로 구분한다. 이 과정에서 위의 툴박스는 n개의 토픽에 대한 각 단어의 빈도수와 분포도, 각의 문서에 대한 토픽의 확률분포도, 각 토픽에 대한 각 문서의 확률분포도를 생성한다.

〈표 3〉은 MA 카테고리의 학습용 데이터세트에 대한 LDA 모델링의 결과로 생성된 45개의 토픽 모델 중 처음 세 토픽의 예이다. 토픽0의 경우 'word'라는 단어의 빈도수가 3,078회, 'text'라는 단어의 빈도수가 2,566회 출현하였다. 이와 같이 표현된 단어와 단어의 빈도수를 바탕으로 토픽0이 주로 문서편집과 관련된 토픽임을 짐작할 수 있다. 마찬가지로 토픽2는 컴퓨터 시스템과 관련한 토픽임을, 토픽3은 학교에서의 발표 슬라이드와 관련한 토픽임을 짐작할 수 있다.



〈그림 2〉 문서-토픽의 확률 분포 - 문서3707의 예

〈표 3〉 LDA 토픽 모델의 예 - Media Arts 카테고리

	토픽0	토픽1	토픽2
상위 10개 단어	Word text document office character [oov] editor format type microsoft	server network share machine client access remote port set ftp	school student kid presentation year slide powerpoint parent people univers

1) http://nlp.stanford.edu/software/tmt/tmt-0.4/

4.2 문서-토픽 확률 분포

실험에 사용한 툴박스의 LDA 모듈을 통해 생성되는 두번째 결과물은 문서-토픽의 확률 분포이다. 〈그림 2〉는 MA 카테고리에 속한 문서3707에 대한 토픽별 확률 분포도이다. 문서3707은 전체 45개의 토픽 중에서 토픽 19, 토픽 24, 토픽 17, 토픽 35, 토픽 12, 토픽 11의 순으로 각각의 토픽을 다루고 있는 것으로 파악되었다. 각 문서가 다루고 있는 토픽의 평균 갯수는 6~7개로 파악되었으며 가장 많은 토픽을 다루고 있는 문서는 최대 23개까지의 토픽을 다루고 있는 것으로 파악되었다.

〈표 4〉 문서별 토픽의 갯수

	CI	MA
Mean	6.2477	7.4622
Max	23	21
Min	1	1

4.3 토픽-문서 확률 분포

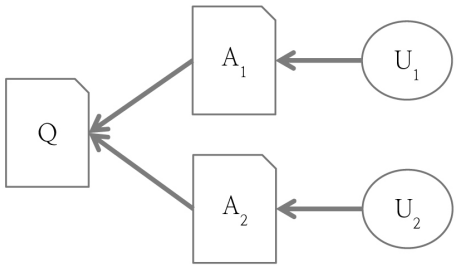
실험에 사용한 툴박스는 LDA모듈을 통해 토픽을 분석한 후 생성하는 세번째 결과물은 토픽-문서의 확률 분포이다. 이 결과에 의하면 MA 카테고리의 경우 11번 토픽은 문서3707에 대해 0.014의 확률로 해당 토픽을 다루고 있음을 파악할 수 있다.

11	3707	0.014354066985645
12	3707	0.071770334928229
17	3707	0.177033492822966
19	3707	0.320574162679425
24	3707	0.267942583732057
35	3707	0.148325358851674

〈그림 3〉 토픽-문서 확률 분포의 예 - MA 카테고리

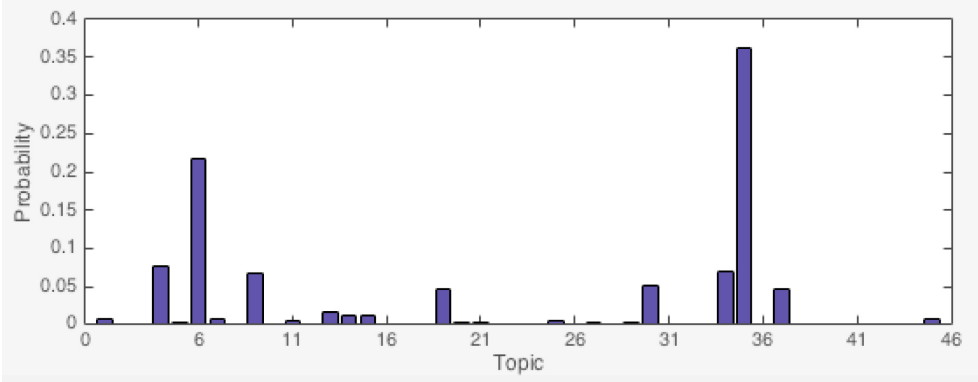
4.4 이용자의 관심 토픽 모델링

본 연구의 궁극적인 목표는 커뮤니티 기반 질의응답 서비스의 환경 하에서 해당 커뮤니티 내에 축적된 질의응답의 코퍼스를 활용하여 이용자의 관심 토픽 분야를 모델링하는 것이다. 따라서 본 연구에서는 이용자의 관심 토픽 분야를 모델링하기 위하여 LDA를 통해 생성된 문서-토픽 확률 분포값을 활용하고자 하였다. 이는 각 이용자가 응답한 질문 또는 답변을 통해 해당 이용자가 응답한 질문 또는 답변의 토픽에 관심을 보인 것으로 가정하고, 이를 통해 해당 이용자의 관심 토픽 분야를 추정하고자 하였다. 본 연구에서는 사용된 문서는 질문자에 의해 커뮤니티에 포스팅된 질문(Q)와 이에 각 이용자(U1, U2)에 의해 제공된 답변으로 구성된 것이다. 따라서, 각 이용자가 답변의 형태로 보여준 행태를 해당 이용자가 관련된 질문 또는 그 질문이 내포하고 있는 토픽에 관심을 보인 것으로 추정할 수 있다.



〈그림 4〉 질문(Q)과 답변(A), 이용자(U)의 관계도

따라서, 각 이용자의 관심 토픽은 각 이용자와 관련있는 혹은 각 이용자가 응답한 질의 및 응답(Q와 A)의 토픽 분포(θ)를 각 이용자 U의 관심 토픽(θu)의 분포로 이해할 수 있다.



〈그림 5〉 평가용 질의의 토픽 분포 - 질의120703의 예

이는 다음과 같은 공식으로 표현할 수 있다.

$$\theta_u \approx \sum_{d=1}^{n} \theta * a(d,u) \qquad \text{(공식1)}$$

여기에서 $a(d, u)$는 이용자 u와 결합된 문서를 의미한다. 결합의 의미는 코퍼스 내에서 이용자가 응답한 질의응답 문서를 의미한다.

$$a(d,u) = \begin{cases} 0 \\ 1 \end{cases}$$

여기서, a의 값은 이용자와 해당 문서가 결합할 경우 1, 결합하지 않는 경우 0의 값으로 계산한다.

이용자의 관심 토픽 분포 값을 산출하기 위한 공식1의 방법은 커뮤니티 내에서 질문을 많이 한 이용자의 토픽 분포에 있어서 그렇지 아니한 이용자에 비해 높은 값을 계산해 낸다. 본 연구에서는 이용자의 관심 토픽을 더욱 잘 표현하는 모델을 찾고자 하므로 공식 1의 방법 이외에도 다음 공식2의 방법을 별도로 시험하였다. 방법2는 커뮤니티 내에서 질의응답을 많이 한 이용자와 그렇지 않은 이용자 사이에 차이를 두지 않고 각 이용자의 순수한 토픽 관심도만 살펴보고자 한 것이다. 이를 위해 다음의 공식으로 이용자의 관심 토픽 분포도를 계산하였다.

$$\theta_u \approx \sum_{d=1}^{n} \theta * a(d,u) * 1/\sum a(d,u) \qquad \text{(공식2)}$$

즉, 공식2는 각 이용자와 결합한 전체 문서의 토픽 분포를 해당 이용자와 결합한 문서의 수로 나눔으로써 커뮤니티 내에서 질의응답을 많이 한 이용자와 그렇지 않은 이용자의 토픽 분포를 정규화함으로써 질의응답 횟수의 차이에 따른 두 이용자 간의 변별력을 없애고자 하였다.

4.5 평가용 질의의 토픽 분석

본 연구에서 실험을 검증하기 위하여 사전에 코퍼스를 학습용 데이터세트와 평가용 데이터세트로 나누어 두었다. 이용자의 관심 토픽 모델을 검증하기 위하여 평가용 데이터세트를 사

용하였다. 평가용 데이터세트는 〈그림 4〉의 문서들(Q, A) 중에서 질의에 해당하는 문서(Q)만을 포함시키고 연관된 응답문서들(A)은 모두 배제하였다. 이는 새로운 질의가 커뮤니티에 유입되는 시점에는 다른 이용자의 답변이 제공되지 않기 때문이다.

일단 새로운 질의가 커뮤니티에 유입되면 이 새로운 질의에 대해 토픽 분석을 실시하여야 한다. 이때 역시 본 연구에서 사용한 툴박스를 이용하여 학습용 데이터세트를 이용하여 생성된 〈표 3〉과 같은 토픽 모델을 대상으로 해당 질의의 토픽 분석을 실시하였다. 이를 통해 각 문서의 토픽 분포 확률을 계산해 낸 것처럼 새로운 질의에 대한 토픽 분포의 확률을 계산해 내었다. 따라서, 새롭게 유입된 질의는 기존의 LDA모델을 바탕으로 여러 개의 토픽을 다루고 있는 것으로 간주된다. 따라서 각각의 질의는 여러 개의 토픽에 대한 확률 분포를 갖게 된다. 〈그림 5〉의 예는 질의120703의 토픽 분포를 나타낸다. 질의120703은 총 21개 토픽을 다루고 있는 것으로 파악되었다. 평가에 사용한 질의의 평균 토픽 갯수는 16개로 학습용의 질의응답 문서가 갖는 평균 6~7개의 토픽 수보다 훨씬 높게 나타났다. 이는 질의문에 포함된 단어의 수가 질의응답 문서에 포함된 단어의 수보다 훨씬 적기 때문에 질의문에 포함된 각 단어가 해당 단어를 포함하는 토픽을 확정할 확률이 그만큼 낮아지기 때문인 것으로 분석된다. 실제로 평가용 질의의 토픽 분포를 이용한 이용자의 후보군의 선정에 있어서 모든 토픽의 확률을 모두 고려한 모델의 경우 최종 평가 결과가 상위 1개의 토픽만을 대상으로 후보를 선정한 모델에 비해 나쁜 결과를 보였다.

〈표 5〉 평가용 질의별 토픽의 갯수

카테고리	CI	MA
Mean	16.21	16.88
Max	32	33
Min	1	1

4.6 이용자 후보군 선정

본 연구의 목적은 커뮤니티에 새롭게 유입되는 질의에 대해 답변을 제공할 것으로 예상되는 이용자를 식별해 내는 데 있다. 위에서 언급한 공식들을 이용하여 이용자의 후보군을 다양한 방법으로 선정할 수 있는데, 본 연구에서는 다음의 방법들을 실험하여 보았다.

4.6.1 이용자의 관심 토픽 점수: Sum vs. Mean

이용자의 관심 토픽 점수를 계산하기 위한 두 가지 공식을 위의 4.4섹션에서 제시하였다. 공식1은 합산(sum)의 방식으로 각 토픽별로 해당 이용자와 결합한 모든 문서의 각 토픽별 확률을 모두 합하는 방식으로, 이용자의 각 토픽별 점수를 계산하였다. 이 방식은 이용자가 특정 토픽에 대해 더 많은 질문을 하면 할수록 더 높은 값을 가지게 되어 결과적으로 많은 질의와 응답을 제공한 이용자를 가중하는 결과를 낳는다. 이에 비해 공식2는 평균값(mean)을 이용하는 방식으로 이용자와 결합한 모든 문서의 각 토픽별 확률을 모두 합한 다음 이를 전체 결합 문서의 수로 나눔으로써 질의응답수에 대한 가중치를 완전히 배제하였다. 이 두가지 공식은 결과적으로 서로 각 이용자에 대해 서로 다른 순위를 제공하여 서로 다른 후보군을 제시하였다.

4.6.2 질의의 토픽수에 따른 실험

4.4에서 설명한 바와 같이 평가용 데이터세트에 포함된 질의문을 대상으로 토픽 분석을 실시한 결과 각 질의가 갖고 있는 토픽의 갯수가 평균 16개로 매우 높게 나타났다. 따라서 이러한 토픽의 갯수가 후보를 선정하는 데 미치는 영향을 알아보기 위하여 해당 질의가 다루고 있는 모든 토픽을 대상으로 후보를 선정하는 방식과 상위 1개의 토픽만을 대상으로 후보를 선정하는 경우를 비교하여 보았다.

4.6.3 후보군 선정의 규모

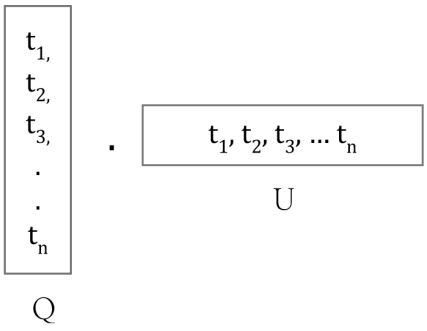
질의가 다루고 있는 토픽에 대해 각 토픽에 대해서 관심을 가지고 있는 이용자를 불러와 이들을 대상으로 후보군을 선정하여야 한다. 본 연구에서는 각 토픽별로 상위 100명의 후보를 대상으로 후보군 선정하는 경우와 상위 500명의 후보를 대상으로 후보군을 선정하는 두 경우를 실험하여 결과를 비교하여 보았다.

4.7 평가용 질의에 대한 이용자의 토픽 점수 계산

커뮤니티에 새롭게 유입된 질의에 대한 각 이용자의 관심 토픽 점수를 계산하기 위하여 다음의 공식을 적용하였다.

$$Score = \sum_{1}^{n} ts_q * ts_u \qquad \text{(공식3)}$$

이용자의 관심 토픽 점수(Score)는 질의에 포함된 각각의 토픽에 대한 확률값(ts_q)에 이용자의 해당 토픽에 대한 관심 토픽 점수(ts_u)를 곱하여 합산한 점수이다. 이 점수의 순위에 따라 이용자를 추천하였다.



〈그림 6〉 질의의 토픽 분포(벡터) * 이용자의 토픽 분포(벡터)

$$Score = \sum_{top1}^{n} ts_q * ts_u \qquad \text{(공식4)}$$

공식4는 질의의 토픽 수를 상위 n개로 한정하여 이용자의 관심 토픽 점수를 계산할 경우에 적용한다. 이는 실험에서 평가용 질의들이 평균 16개의 토픽을 가지고 있어 각 질의가 다루고 있는 토픽 분석의 결과가 정확하지 않을 수 있어 다루고 있을 토픽의 확률이 높은 순으로 상위 n개의 토픽을 한정하여 결과값을 산출하고자 할 경우에 적용하였다. 본 연구에서는 최상위 토픽을 대상으로 한 경우와 모든 토픽을 대상으로 한 경우의 차이를 보기 위해여 n의 값을 1로 정하여 실험하였다.

4.8 적합성 판단

본 연구의 주요 과제는 평가용 질의에 대해 답변을 제공할 것으로 예측되는 이용자를 각 이용자의 관심 토픽 모델을 바탕으로 추천하는 것이다. 따라서, 위에서 제시한 여러 모델을 통하여 다양한 순위의 이용자를 제공할 경우, 제공

된 이용자가 적합한 지 그렇지 않은 지를 평가해야 한다. 이 적합성의 판단은 실제 평가용 데이터세트에서 평가용 질의에 응답한 이용자들을 적합한 이용자로 간주하여 이 연구에서 제시하는 각 모델의 성능을 평가하였다.

4.9 평가 지표

각각의 모델이 어느 정도의 성능을 보이는지 평가하기 위하여 Mean Average Precision(MAP)을 측정하여 각 모델들 간의 성능을 평가하였다. 이는 평가용 데이터세트 내의 질의의 수가 많기 때문에 전체 질의에 대한 모델의 정확성을 평가하고자 MAP을 측정하였다.

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}$$

즉, MAP은 각 질의에 대한 평균 정확율(AP)를 계산하여 이를 다시 전체의 질의 수로 나누어 평균 값을 구한 것이다.

$$AP = \frac{\sum_{k=1}^{n} rel(k)}{\#relevant\ documents}$$

여기에서 rel(k)는 k번째 랭크된 적합문서의 값으로 적합 문서의 출현시 1로 계산하고 그렇지 않은 경우 0으로 계산한다.

5. 결과 분석

본 연구에서 실험한 각각의 모델에 대한 성능을 평가하기 위하여 각 모델과 상대 모델에 대해 MAP을 계산하여 비교하였다. 또한 각각의 모델에 대해 동일한 평가용 데이터 세트를 사용하였기 때문에 각 모델에 따른 변화가 유의미한 지를 파악하기 위하여 paired t-test 기법을 통하여 검증하였다.

5.1 이용자의 관심 토픽 점수 계산 방법에 따른 차이: Sum vs. Mean

〈표 6〉은 이용자의 관심 토픽 점수를 계산하는 방법의 차이에 따른 성능평가의 결과값이다. CI와 MA 두 카테고리 모두의 경우에 있어서 Sum의 방식으로 이용자의 관심 토픽 점수를 계산한 방식(Sum모델, 공식1 적용)이 이용자의 관심 토픽 점수를 정규화한 방식(Mean모델, 공식2 적용)보다 훨씬 나은 결과를 보였다. 이 두 모델의 성능에 있어서의 차이가 유의미한 지를 알아보기 위해 t-test 검증을 실시하였다. 검증결과, CI카테고리의 경우, Sum모델(M=0.023, SD=0.059)과 Mean모델(M=0.000, SD=0.001)의 MAP 점수에 있어서 유의미한 차이가 있음을 확인하였다; t(27.89), df=5140, p=0.000. 마찬가지로, MA카테고리에서도 Sum모델(M=0.019, SD=0.061)과 Mean모델(M=0.000, SD=0.000)의 MAP 점수에 있어서 유의미한 차이가 있음을 확인하였다; t(16.07), df=2591, p=0.000. 이는 커뮤니티 내에서 이용자의 많은 질의응답 활동이 이용자의 관심 토픽을 추정하는 데 상당히 의미있는 영향을 미치는 것임을 알 수 있다. 따라서, 커뮤니티 기반 질의응답 서비스를 위한 이용자의 관심 토픽을 모델링 하는 데 있어서, 이용자의 커뮤니티 내에서의 질

의응답 활동의 정도를 고려하는 것이 더 좋은 결과를 낼 수 있을 것으로 생각된다.

〈표 6〉 이용자의 관심 토픽 점수 계산 방식에 따른 MAP@500의 차이

	CI	MA
Sum	0.023174*	0.019305*
Mean	0.000069	0.000043

* 유의미한 차이가 있음.

5.2 질의의 토픽 수에 따른 차이

질의에 포함된 토픽 수에 따른 성능의 차이를 보기 위해 모든 토픽을 대상으로 이용자의 후보를 선정하는 방식(TopicAll 모델)과 가장 높은 확률을 지닌 토픽만을 대상으로 이용자의 후보를 선정하는 방식(TopicTop 모델)을 비교하여 보았다. 이 경우 〈표 7〉에서와 같이 최상위 토픽을 대상으로 한 TopicTop 모델이 모든 토픽을 대상으로 한 TopicAll모델보다 주어진 질의에 대해 응답한 이용자를 예측하는 데 있어서 향상된 결과를 보였다. 주어진 질의에 대한 토픽을 분석한 후, 대상 토픽을 선정하는 데 있어서 서로 다른 토픽 수가 어떠한 차이를 보이는지 알아보기 위하여 위의 두 모델에 대해 t-test 검증을 실시하였다. 검증결과, CI카테고리에서는 TopicTop모델(M=0.025, SD=0.066)과 TopicAll모델(M=0.023, SD=0.059)의 MAP 점수에 있어서 유의미한 차이가 있음을 확인하였다; t(4.04), df=5140, p=0.000. 마찬가지로, MA카테고리에서도 TopicTop모델(M=0.021, SD=0.062)과 TopicAll모델(M=0.019, SD=0.061)의 MAP 점수에 있어서 유의미한 차이가 있음을 확인하였다; t(2.69), df=2591, p=0.007. 이는 커뮤니티 기반 질의응답 서비스 환경에서 주어진 질의에 대한 토픽 분석과 그에 따른 후보군의 선정시 질의에 내포된 토픽 중에서 상대적으로 높은 확률을 지닌 토픽을 대상으로 이용자를 선정하는 것이 더 나은 결과를 나타낼 것으로 예상된다. 이를 위한 한 방법으로 주어진 질의에 대해 적절한 수의 토픽 수를 정하기 위한 토픽 분포의 확률 값의 하한선(threshold)을 기계학습 등의 방법을 통해 정하는 것이 한 방법이 될 수 있겠다.

〈표 7〉 질의의 토픽수에 따른 MAP@500의 차이

토픽	CI	MA
TopicAll	0.023174	0.019305
TopicTop	0.025187*	0.020606*

* 유의미한 차이가 있음.

5.3 후보군의 선정의 규모에 따른 차이

실제 커뮤니티 기반 질의응답 서비스에서 새로운 질의가 유입될 때 이에 대해 모든 이용자를 대상으로 주어진 질의의 토픽 분포와 이용자의 토픽 분포를 모두 비교하는 것은 현실적으로 불가능하다. 본 연구에서는 새로 유입된 질의의 토픽을 분석한 후 각 토픽에 대해 해당 토픽과 결합한 이용자들 중에서 각 이용자의 해당 토픽에 대한 관심 토픽의 점수에 따라 상위 100명 또는 500명의 후보군을 선정하여 이들을 대상으로 이용자의 토픽 분포에 대한 점수를 계산하였다(공식4).

〈표 8〉에서와 같이 후보군의 규모를 500명으로 증가시킨 경우(TOP500모델), 상대인 100명의 후보군을 각 토픽과 결합한 이용자들로부

터 가져오는 경우(TOP100모델)에 비해 MAP의 점수에 증가를 보였다. 이러한 MAP점수의 변화의 유의미성을 검증하기 위해 두 모델에 대해 t-test검증을 실시하였다. 검증결과, CI 카테고리에서는TOP100모델(M=0.023, SD=0.065)과 TOP500모델(M=0.025, SD=0.066)의 MAP 점수에 있어서 유의미한 차이가 있음을 확인하였다; t(-51.26), df=5140, p=0.000. MA카테고리에서도 TOP100모델(M=0.019, SD=0.062)과 TOP500모델(M=0.021, SD=0.062)의 MAP 점수에 있어서 유의미한 차이가 있음을 확인하였다; t(-37.92), df=2591, p=0.000. 이는 특정 토픽과 결합한 이용자들 중에서 너무 적은 수의 이용자를 가져올 경우, 그 후보군 중에서 실제 적합한 이용자가 제외될 가능성이 높아지기 때문인 것으로 분석된다. 즉, 최종 후보군을 선정하기 전, 예비 단계로 각 토픽과 결합한 후보 이용자군을 불러올 때에는 적합한 이용자가 이 단계에서 누락되지 않도록 적절한 수준으로 후보 이용자군의 규모를 조절하는 것이 필요해 보인다.

〈표 8〉 후보군 선정의 규모에 따른 MAP의 차이 - Top1 모델

	CI	MA
TOP100	0.023433	0.018759
TOP500	0.025187*	0.020606*

* 유의미한 차이가 있음.

6. 결 론

최근 인기가 증가하고 있는 커뮤니티 기반의 질의응답 서비스는 이용자들간에 서로 묻고 답하는 형식으로 서로 도움을 제공하여 성공적으로 발전하고 있다. 최근에는 이용자의 수가 급증한 것과 비례해 해당 커뮤니티에 매일 새롭게 유입되는 질의가 상당하다. 이러한 질의를 아무런 시스템의 도움 없이 이용자들에게 맡겨두어 해결하도록 하는 데 많은 어려움을 겪어왔다. 이러한 문제를 해결하는데 있어서 질의할당(question routing)이 어느 정도 도움을 제공할 것이라고 생각한다.

본 연구에서는 질의할당을 하기 위해 기존의 커뮤니티 기반 질의응답 사이트에 축적된 질의응답의 코퍼스를 이용하여 해당 커뮤니티 내의 각 이용자의 토픽 관심도를 측정하고 이를 바탕으로 새로 유입되는 질의의 토픽을 분석한 후 적합한 이용자에게로 해당 질의를 할당하는 방식에 대해 연구해 보았다. 주어진 질의에 대해 응답할 가능성이 높은 이용자들을 식별하기 위하여, 이용자의 토픽을 분석하기 위한 한 방법으로 이용자와 결합한 문서들을 대상으로 LDA를 이용해 토픽분석을 실시한 후 해당 문서가 다루고 있을 것으로 분석된 각각의 토픽들을 해당 이용자의 관심 토픽으로 추정하여 이용자의 관심 토픽을 모델링하고자 시도하였다. 이러한 접근 방법의 장점은 토픽 분석에 널리 알려진 LDA 기법을 적용하여 쉽게 각 문서에 내포되어 있는 토픽을 분석할 수 있는 데 있다. 문서 내에 숨겨져있는 여러 토픽을 분석해 내는 LDA기법은 비슷한 토픽을 다루고 있는 문서들을 분류하는 데 주로 사용되고 있는데, 본 연구에서는 이를 비슷한 토픽에 대한 관심을 가진 이용자들을 식별해 내는 데 활용한 데 의의가 있다.

본 연구에서 이용자의 관심 토픽을 모델링하

기 위하여 LDA기법을 사용하였는데, LDA는 하나의 문서 속에 여러 개의 토픽을 다루고 있다고 간주하므로, 이러한 유연성이 이용자의 관심 토픽을 단 하나로 제한하지 않아 이용자의 관심 토픽을 모델링하는 데 적합하다고 생각된다. 그러나 연습용 데이터세트와 평가용 데이터세트를 대상으로 각 질의에 대한 토픽 분석을 시도한 결과 이용자들의 답변이 포함되지 않은 질의만으로 구성된 데이터세트에 대한 토픽 분석의 결과가 연습용 데이터세트의 결과에 비해 평균 두 배나 많은 토픽의 수를 예측해 내었다. 이는 연습용 데이터세트를 통해 생성된 토픽 모델이 평가용 데이터세트에 일정부분만 기능하는 것으로 판단된다. 본 연구의 결과에 가장 큰 영향을 미치는 부분은 이용자의 관심 토픽 모델링 뿐만 아니라, 새로 유입되는 질의에 대한 토픽을 보다 정확히 분석해 내는 것이다. 향후, 이를 개선하기 위해 Ask Metafilter의 질의응답 코퍼스에서 제공되고 있는 것처럼 이용자가 자신의 질의에 대해 미리 정의한 태그들을 활용하여 평가용 질의에 대한 토픽 분석의 수준을 향상시킬 필요가 있겠다. 본 연구에서 유의미하게 관찰된 점은 커뮤니티 내에서 보다 적극적으로 질의응답을 제공한 이용자의 활동이 주어진 질의에 대해 적합한 이용자를 예측하는 데 상당히 긍정적인 영향을 미치는 점이다. 따라서, 커뮤니티에 기반한 질의응답 서비스에서 주어진 어떤 질의 혹은 토픽에 대해 전문성 또는 관심을 가진 이용자를 탐색하는데 이용자의 질의응답 활동을 적절하게 활용할 필요가 있겠다. 본 연구의 제약점은 주어진 질의에 대해 응답할 가능성이 있는 이용자를 예측하기 위해 여러 모델들을 실험하여 유의미한 성능의 향상을 관찰하였지만, 여전히 전체적으로 매우 낮은 정보검색 성능을 보이고 있다. 이는 연구에 커뮤니티 기반 질의응답 서비스의 특징 중의 하나인 전체 이용자 가운데 극소수의 이용자만이 질문에 답변하는 경향으로 인해 학습 및 평가에 사용된 질의 중 상당 부분이 매우 적은 수의 적합한 답변자를 가지고 있기 때문으로 추측된다. 따라서, 향후 연구에서는 커뮤니티 기반 질의응답 서비스에서 이용자들의 응답 행태에 대한 보다 면밀한 분석이 요구된다. 또한 적합 이용자의 희소성의 문제를 극복하기 위하여 해당 사이트로부터 수집된 데이터를 바탕으로 보다 효율적인 특성들을 찾아내고 이를 종합적으로 활용하여 질의할당을 위한 이용자를 추천할 필요가 있겠다.

참 고 문 헌

Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008, April). Knowledge sharing and yahoo answers: everyone knows something. In Proceedings of the 17th international conference on World Wide Web (pp. 665-674). ACM.

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008, February). Finding high-quality content in social media. In Proceedings of the 2008 International Conference on Web Search

and Data Mining (pp. 183-194). ACM.

Bouguessa, M., Dumoulin, B., & Wang, S. (2008, August). Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 866-874). ACM.

Chang, S., & Pal, A. (2013, August). Routing questions for collaborative answering in community question answering. In Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on (pp. 494-501). IEEE.

Farzan, R., & Brusilovsky, P. (2011). Encouraging user participation in a course recommender system: An impact on user behavior. Computers in Human Behavior, 27(1), 276-284.

Jeon, J., Croft, W. B., & Lee, J. H. (2005, October). Finding similar questions in large question and answer archives. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 84-90). ACM.

Jurczyk, P., & Agichtein, E. (2007). Discovering authorities in question answer communities by using link analysis. Paper presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.

Kim, Soojung (2012). Research Trends of the Credibility of Information in Social Q&A. Journal of the Korean Society for Information Management, 29(2), 135-154.

Liu, M., Liu, Y., & Yang, Q. (2010). Predicting best answerers for new questions in community question answering. In Web-Age Information Management (pp. 127-138). Springer Berlin Heidelberg.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems, 2, 849-856.

Oh, S. (2010). Answerers' Motivations and Strategies for Providing Information and Social Support in Social Q&A an Investigation of Health Question Answering. ProQuest LLC. 789 East Eisenhower Parkway, PO Box 1346, Ann Arbor, MI 48106.

Qu, M., Qiu, G., He, X., Zhang, C., Wu, H., Bu, J., & Chen, C. (2009, April). Probabilistic question recommendation for question answering communities. In Proceedings of the 18th international conference on World wide web (pp. 1229-1230). ACM.

Shah, C. (2011). Measuring effectiveness and user satisfaction in Yahoo! Answers. First Monday, 16(2).

Zhang, J., Tang, J., & Li, J. (2007). Expert finding in a social network. In Advances in Databases: Concepts, Systems and Applications (pp. 1066-1069). Springer Berlin Heidelberg.