QualityRank : 소셜 네트워크 분석을 통한 Q&A 커뮤니티에서 답변의 신뢰 수준 측정

(QualityRank : Measuring Authority of Answer in Q&A Community using Social Network Analysis)

김 덕 주 † 박 건 우 †† 이 상 훈 †††
(DeokJu Kim) (GunWoo Park) (SangHoon Lee)

요 약 질문(Question)과 답변(Answer)을 하는 커뮤니티 기반의 지식검색서비스에서는 질의를 통해 원하는 답변을 얻을 수 있지만, 수많은 사용자들이 참여함에 따라 방대한 문서 속에서 신뢰성있는 문서를 찾아내는 것은 점점 더 어려워지고 있다. 지식검색서비스에서 기존 연구는 사용자들이 생성한 데이터 즉 추천수, 조회수 등의 비텍스트 정보를 이용하거나 답변의 길이, 자료첨부, 연결어 등의 텍스트 정보 이용하여 문서의 품질을 평가하고, 이를 검색에 반영하여 검색성능을 향상시키는 데 활용했다. 그러나 비텍스트 정보는 질의/응답의 초기에 사용자들에 의해 충분한 정보를 확보할 수 없는 단점이 있으며, 텍스트 정보는 전체의 문서를 답변의 길이, 연결어등과 같은 일부요인으로 판단해야하기 때문에 품질평가의 한계가 있다고 볼 수 있다. 본 논문에서는 이러한 비텍스트 정보와 텍스트 정보의 문제점을 개선하기 위한 QualityRank 알고리즘을 제안한다. QualityRank는 텍스트/비텍스트 정보와 소셜 네트워크 분석 기반의 사용자 중앙성을 고려하여 질문에 적합하고 신뢰성 있는 답변을 랭킹화 한다. 실험결과 제안한 알고리즘을 사용했을 경우 텍스트/비텍스트 모델 보다 랭킹성능에 있어 향상된 결과를 얻을 수 있었다.

키워드 : 지식검색서비스, QualityRank 알고리즘, 소셜 네트워크 분석

Abstract We can get answers we want to know via questioning in Knowledge Search Service (KSS) based on Q&A Community. However, it is getting more difficult to find credible documents in enormous documents, since many anonymous users regardless of credibility are participate in answering on the question. In previous works in KSS, researchers evaluated the quality of documents based on textual information, e.g. recommendation count, click count and non-textual information, e.g. answer length, attached data, conjunction count. Then, the evaluation results are used for enhancing search performance. However, the non-textual information has a problem that it is difficult to get enough information by users in the early stage of Q&A. The textual information also has a limitation for evaluating quality because of judgement by partial factors such as answer length, conjunction counts. In this paper, we propose the QualityRank algorithm to improve the problem by textual and non-textual information. This algorithm ranks the relevant and credible answers by considering textual/non-textual information and user centrality based on Social Network Analysis(SNA). Based on experimental validation we can confirm that the results by our algorithm is improved than those of textual/non-textual in terms of ranking performance.

Key words : Knowledge Search service, QualityRank algorithm, Social Network Analysis

† 학생회원 : 국방대학교 전산정보학과
cocobi1@hanmail.net
†† 정 회 원 : 국방대학교 전산정보학과
pgw4050@emerald.yonsei.ac.kr
††† 종신회원 : 국방대학교 전산정보학과 교수
07uandme@gmail.com
논문접수 : 2010년 8월 23일
심사완료 : 2010년 9월 28일

정보과학회논문지: 데이타베이스 제37권 제6호(2010.12)

1. 서 론

지식검색 서비스는 사용자가 자발적으로 참여하여 상호 질문과 답변을 하는 커뮤니티 기반 서비스로 네이버 지식iN을 시작으로 다음(daum), 야후(yahoo) 등과 같은 포탈들의 참여로 국내 검색의 대표적 서비스로 성장하였다. 이러한 추세는 누구나 어떠한 주제에 대해서도 질문과 답변을 할 수 있다는 개방적 구조와 이렇게 축적된 자료를 공유할 수 있다는 커뮤니티적인 특성에 기인한다. 하지만 불특정 다수의 사용자에 의해 구축된 방대한 자료 속에는 검증되지 않은 답변이나 추측성 답변들로 인하여 답변의 신뢰성, 정확성, 전문성이 저하되고 지식의 질적 하락이 초래될 수 있는 문제점이 있다. 이러한 문제점은 사용자가 진정으로 원하는 답변 획득을 점점 더 어렵게 만들고 있다. 이를 해결하기 위해 지식검색 서비스의 결과물로 대변되는 답변 문서의 특성을 평가하기 위한 기준을 제시한 연구가 수행되었다[1]. 연결어, 추정어, 개인의견, 광고성 단어, 가치 판단어, 이모티콘 등의 텍스트 요소로 문서의 신뢰도를 평가하는 연구[2]와 문서의 조회 수나 추천 수 등의 비텍스트 정보를 이용하는 연구도 수행되었다[3]. 하지만 비텍스트 정보는 질의/응답의 초기 사용자들에 의해 충분한 정보를 제공할 수 없다는 단점과 텍스트 정보는 전체의 문서를 연결어, 추정어, 가치 판단어 이모티콘 등의 일부요인으로 판단해야 하는 문제점을 가지고 있다.

본 논문은 지식검색 서비스의 이러한 문제점을 텍스트 요소와 비텍스트 정보, 소셜 네트워크 관점에서 해결하고자 한다. 이는 객관적 수치를 바탕으로 높은 품질 지수를 갖는 양질의 질문/답변을 통해 지식공유의 근본적 목적에 부합될 수 있을 것이라는 가정에서 시작되며, 이를 위해 대표적인 국내 지식검색 서비스인 네이버 지식iN에서의 텍스트 정보와 비텍스트 정보를 바탕으로 카테고리별 Quality Rank 알고리즘을 제안한다.

논문은 2장에서는 관련연구, 3장에서는 소셜 네트워크 및 텍스트/비텍스트 기반 요소 추출 및 이를 통한 QualityRank 알고리즘을 제안한다. 4장에서는 실험을 위한 데이터 셋과 제안한 알고리즘을 통한 품질 지수 산정 및 평가 결과를 제시하고 마지막으로 결론 및 향후 연구 과제를 제시한다.

2. 관련연구

2.1 집단지성과 지식검색

지식과 정보의 구분이 모호해 지면서 등장한 지식의 여러 유형 가운데 집단 지성은 온라인이란 환경에 힘입어 큰 영향력을 미치고 있다. 이러한 지식은 광의의 개념으로서 일상생활과 관련된 지식, 다양한 지식 생산자가 제공하는 지식, 상대적으로 불안정하고 유동적인 지식, 집합적으로 구성되는 지식이다[4]. 즉, 온라인에서의 지식은 우리의 일상생활과 관련된 정보, 상식, 조언까지도 포함하는 보다 확장된 개념으로서, 사용자를 포함한 다양한 지식 생산자들이 직접 제공하는 상대적으로 불안정하고 유동적인 지식이다. 또한 이를 기반으로 한 지식검색 시스템에서 제공하는 지식은 현재 나의 목적에 어떠한 의미가 있는가에 따라 현재 시점에서 창출되는 지식으로 인터넷에서 집합적으로 공유되고 끊임없이 구성되는 특성을 지닌다. 이러한 점은 집단지성의 발현과 관련이 깊다. 집단지성은 다수의 사용자가 개개인의 작업 및 지식을 공유하고 취합하여 일반적 사실을 도출해 낸다는 원리를 가지고 있다. 이 원리는 다수 사용자의 참여에 의해 어떤 사실에 대한 해결의 실마리를 얻는다는 것이 핵심이다[5].

2.2 소셜 네트워크 분석

소셜 네트워크는 최근 온라인을 중심으로 하여 하나 이상의 상호 의존적인 관계에 의해 구성된 개인 또는 집단의 사회적 구조체(Social Structure)로 정의할 수 있다. 대표적인 소셜 네트워크 서비스(SNS : Social Network Service)인 위키피디아, 트위터(twitter), 페이스북(facebook), 한국의 싸이월드 등에서 볼 수 있듯이 정보과학 분야에서는 기 구성된 소셜 네트워크의 현상을 웹 환경에 응용하는 연구가 활발히 진행 중이다. 이러한 연구는 첫째, 소셜 네트워크상에 존재하는 웹 사용자 간의 연결성(Connectivity) 확장을 통한 검색 효율의 향상 방법과 둘째, 실제 사회 현상과 소셜 네트워크상의 현상에 대한 비교 분석, 마지막으로 네트워크 구성의 효율성 및 보안 등 소셜 네트워크 자체에 대한 연구로 나눌 수 있다[6]. 사회학, 통신공학, 경제학 등에서 폭넓게 연구 중인 소셜 네트워크 분석은 소셜 네트워크의 형태와 특성을 알고리즘 적으로 연구하는 것으로 전체 관계망에서의 위치와 그 효과를 측정하는 위치적 접근법(Positional approach)과 연결망의 직접적인 관계에 초점을 둔 관계적 접근법(Relational Approach)으로 분류된다[7]. 소셜 네트워크의 분석은 노드간의 관계 구조를 찾아내기 위해 그래프 이론을 이용한 소셔메트리(Sociometry)와 수학적 방법인 계량적 방법을 이용한다. 수학적 방법의 기본은 행렬과 그래프의 이해이다. 구성원 (i, j) 사이의 관계가 있고 없음을 1과 0으로 나타내는 행렬을 인접행렬(Adjacency Matrix)이라고 부르며 행렬의 항(Cell)은 i로부터 j에 이르는 관계로 표현된다. 그림 1은 인접행렬과 각 항의 관계를 네트워크로 표현하였다.

3. 답변문서 품질 측정

3.1 품질지수(Quality Value)

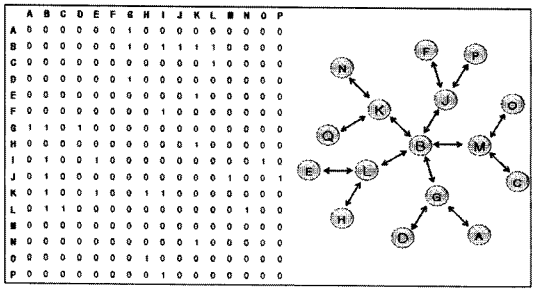


그림 1 인접행렬과 그래프

품질의 사전적 의미는 "물건의 성질과 바탕"이다. 일반적으로 문서 품질은 문서의 성질과 바탕이 되며, 문장의 문법, 문체 등 어법적 관점에서의 정확성과 작성된 내용이 주제에 맞게 적절한지, 작성자의 의도가 명확한지 등으로 설명할 수 있다.

문서의 품질을 평가하기 위한 기준은 주관적인 것이지만, 주어진 질문에 대하여 사용자가 읽고 충분히 신뢰할 만큼 성실하고 자세히 기술된 답변을 높은 품질의 답변이라고 보고, 이에 비해 신뢰하기 어려울 정도로 불성실하고 부실하게 작성한 답변을 낮은 품질의 답변이라고 간주한다. 문서의 내용이 얼마나 성실하고 꼼꼼하게 작성되었고 따라서 그 문서를 읽는 사용자로 하여금 내용에 얼마나 신뢰할 수 있는지를 측정하는 작업은 실제로 매우 어렵고, 관점에 따라 복잡하고 시간을 요하는 과정이 필요하다고 볼 수 있다. 하지만 본 연구에서 품질지수는 "지식검색 서비스에서 질의/답변의 품질을 텍스트 요소와 비텍스트 요소를 고려하여 지수화한 것"으로 웹에서 제공되는 다수의 질의/답변 문서가 품질평가 대상이 된다. 서비스에서 제공되는 추천수, 조회수 등의 비텍스트 요소와 문서 자체의 길이, 답변에 주로 출현하는 단어 등과 같은 텍스트 요소, 질문자와 답변자의 관계를 나타낸 소셜 네트워크 요소를 고려하여 컴퓨터가 계산하기 쉬운 변수들을 통해 품질지수를 도출하였다.

3.2 요소 선정

3.2.1 텍스트 요소

지식검색 서비스에서 답변 내용 자체에 대한 신뢰성을 반영하기 위해 답변 자체 텍스트에 대한 요소를 고려한 것이다. 높은 품질의 답변에서 주로 출현하는 단어와 낮은 품질의 답변에서 자주 출현하는 단어들을 미리 사전으로 작성하고, 사전에 기록된 단어가 답변에 등장한 비율을 측정한다[2].

- 연결어 출현 비율 : 문장의 연결을 위해 사용되는 연결어가 주어진 답변에서 차지하는 비율
- 구체화 단어 출현 비율 : 문장의 이해를 쉽게 구체화하기 위해 사용된 단어가 주어진 답변에서 차지하는 비율
- 멀티미디어자료 출현 비율 : 동영상, 사진, 웹사이트의 주소가 답변에서 차지하는 비율
- 이모티콘의 출현 비율 : 주로 감정을 나타내는 이모티콘이 주어진 답변에서 차지하는 비율
- 답변의 길이 : 답변에서 단어들의 합

표 1 텍스트 요소의 예제

요소	예제
연결어	그리고, 그러나, 그러므로, 따라서 등
구체화 단어	쉽게 말하면, 다시 말하면, 예를 들면 등
멀티미디어	동영상, 사진, 소리, 웹사이트주소 등
이모티콘	^^, --, ㅠㅠ

표 2 텍스트 요소 산정

순번	요소	타입
1	연결어 출현비율	퍼센트
2	구체화 단어 출현비율	퍼센트
3	멀티미디어 출현비율	퍼센트
4	이모티콘 출현비율	퍼센트
5	답변의 길이	정수

표 1은 연결어, 구체화 단어 등 답변의 신뢰도에 영향을 줄 수 있는 단어들을 뽑아 단어목록을 사전으로 작성하였다.

표 2는 표 1와 같이 추출된 연결어, 구체화 단어 등을 텍스트 요소로 산출하였다.

3.2.2 비텍스트 요소

지식검색 서비스는 사용자가 궁금한 것을 질문하고, 질문 기간 동안 다른 사용자가 이것에 대해 답변을 하는 형식으로 이루어진다. 질문자는 질문 기간이 늘어날수록 수 많은 답변들 중 하나의 답변을 채택한다. 사용자들은 해당 질문을 조회, 적절한 답변을 추천/선택하거나 추가 답변을 작성하는 등의 활동을 통해 비텍스트 요소를 생성한다. 조회수, 질의자 등급, 답변추천, 질문자채택, 네티즌채택 등 사용자들에 의해 제공되는 비텍스트 요소를 품질요소로 보고 표 3과 같이 선정하였다[3].

표 3 비텍스트 요소 선정

순번	요소	타입
1	추천수	정수
2	질의자 등급	등급
3	답변수	정수
4	질문자채택	이진수
5	네티즌채택	이진수

3.2.3 소셜 네트워크 요소

질문에 대한 수 많은 답변 중 질문자가 답변을 선택하는 행위는 적절한 답변을 선택하는 것으로 높은 품질의 답변일 가능성을 제시한다. 소셜 네트워크 분석 기법을 적용하여 카테고리별로 질문자와 답변 채택자 관계

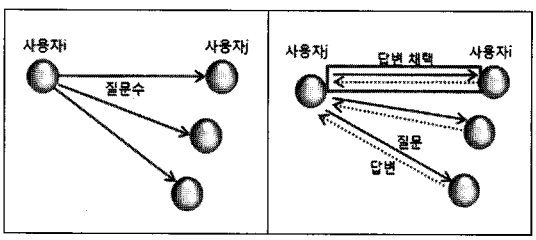
를 나타내는 네트워크를 구축하고 이를 통해 각 사용자의 중앙성 지수를 산출하였다. 한 사용자(노드)가 네트워크 내에서 중앙성을 갖는다는 것은 지식검색 서비스에서 다수의 다른 노드와 연결 관계를 갖는다는 것이다. 이를 사용자간의 지식공유 측면으로 이해한다면 노드를 향해 오는 내향 중앙성(In-Centrality)은 식 (1)과 같고, 밖으로 나가는 외향 중앙성(Out-Centrality)은 식 (2)로 표현된다[8].

$$indegree_{ik} = \sum_{i=1}^{N} Z_{ijk} = Z_{jk} \quad (1)$$

$$outdegree_{ik} = \sum_{j=1}^{N} Z_{ijk} = Z_{ik} \quad (2)$$

• Z_{ijk} : k연결망에서 i구성원으로 부터 j구성원까지의 관계

식 (1), (2)는 일반적인 소셜 네트워크 분석의 내/외향 연결정도에 대한 수식이며 식 (3)은 이를 지식검색 서비스의 사용자에게 적용한 것이다. 한명의 사용자가 질문을 하면 수 많은 다른 사용자가 답변을 하게 된다. 답변은 질의자와 사용자들에 의해 채택이 되는 과정 거친다. 그림 2는 사용자가 질의하고 채택되는 과정을 네트워크화 한 것으로 사용자의 중앙성을 다른 사용자로 가는 질문수(outdegree)와 다른 사용자로 부터 받은 답변 채택수(indegree)로 표현할 수 있다. 이를 기반으로 지식검색 서비스에서 카테고리내 사용자의 중앙성 지수(C_i)를 식 (3)으로 유도하였다.



(a) 질문수(outdegree) (b) 답변 채택수(indegree)

그림 2 사용자 질문답변의 네트워크 구조

$$C_i = \frac{indegree + outdegree}{k-1} \quad (3)$$

indegree : 사용자 i가 다른 사용자 j로부터 받는 답변 채택수

outdegree : 사용자 i로부터 다른 사용자 j에게 가는 질문수

k : 네트워크에 존재하는 전체 사용자수

표 4는 소셜 네트워크 분석을 통해 사용자 중앙성 요소를 선정하였다.

표 4 소셜 네트워크 요소 선정

순번	요소	타입
1	사용자 중앙성	정수

3.3 알고리즘 산출

다수의 지식제공자들이 집합적으로 구성하는 지식검색 서비스에서의 가장 큰 장점은 지식의 생산이라는데 있다. 하지만 수많은 답변들 중에 사용자의 참여는 지식의 전문성 및 신뢰성을 떨어뜨린다. 이에 기존연구는 사용자들이 생성한 데이터 즉 추천수, 조회수 등의 비텍스트 정보를 이용하거나 답변의 길이, 자료첨부, 연결어 등의 텍스트 정보 이용하여 전문가를 식별하거나 문서의 품질을 평가하고, 이를 검색에 반영하여 검색성능을 향상시키는 데 활용했다. 비텍스트 정보는 질의/응답의 초기에 사용자들에 의해 충분한 정보를 확보할 수 없는 문제점이 제기되며, 텍스트 정보는 전체의 문서를 답변의 길이, 자료 첨부 등과 같은 일부요인으로 판단해야하기 때문에 품질평가의 한계가 있다고 볼 수 있겠다. 본 논문에서는 이러한 단점을 보완할 수 있도록 텍스트/비텍스트 특성과 소셜 네트워크의 사용자 중앙성을 기반으로 한 알고리즘을 제안한다.

• 텍스트/비텍스트 품질지수 산정

산정된 텍스트/비텍스트 요소에 대해 최대 엔트로피 모델[9]을 기반으로 문서에 대한 품질지수를 산출할 수 있다. 주어진 답변을 문서 X라고 하고 이 X에 매길수 있는 품질 등급을 y={높음, 보통, 낮음}라 하자. 품질 평가 모델의 목적은 조건부 확률 p=(y=높음|x) 즉, 주어진 문서가 높은 품질의 문서일 확률을 구하는 것이다. 최대 엔트로피 모델을 이용하면 p(y|x)는 식 (4)와 같이 계산된다[3].

$$p(y|x) = \frac{1}{Z}\exp\left[\sum_{i=1}^{k}\lambda_i f_i(x)\right] \quad (4)$$

• $f_i(x)$: i 번째 요소의 값을 출력하는 함수

• λ_i : i 번째 요소의 가중치

• Z : 정규화 상수, k : 총 요소의 개수

식 (4)를 텍스트, 비텍스트 요소에 적용하면 식 (5), (6)이 유도된다.

$$QV_{\text{텍스트}} = \frac{1}{Z}exp\begin{pmatrix}\alpha_1 f_1(x)+\alpha_2 f_2(x)+\alpha_3 f_3(x)\\ +\alpha_4 f_4(x)+\alpha_5 f_5(x)\end{pmatrix} \quad (5)$$

• $f_{1\sim5}(x)$: 선정된 5개의 텍스트 요소를 출력하는 함수

• $\alpha_{1\sim5}$: 텍스트 요소의 가중치

$$QV_{\text{비텍스트}} = \frac{1}{Z}exp\begin{pmatrix}\beta_1 g_1(x)+\beta_2 g_2(x)+\beta_3 g_3(x)\\ +\beta_4 g_4(x)+\beta_5 g_5(x)\end{pmatrix} \quad (6)$$

• $g_{1\sim5}(x)$: 선정된 5개의 텍스트 요소를 출력하는 함수

• $\beta_{1\sim5}$: 비텍스트 요소의 가중치

• QualityRank 알고리즘

제안하는 품질지수 QV(Quality value)는 위와 같이 식 (5), (6)의 텍스트/비텍스트 요소를 고려한 최대 엔트로피 모델과 식(3)의 소셜 네트워크에서 중앙성 지수를 적용하여 식 (7)와 같이 유도하였다.

$$QV = \frac{1}{Z} exp\left[\sum_{i=1}^{m} \alpha_i f_i(x) + \sum_{i=1}^{n} \beta_i g_i(x)\right] \times \gamma\left(\frac{\ln degree_{ij} + outdegree_{ij}}{k-1}\right) \quad (7)$$

• Z : 정규화 상수, $\alpha_i, \beta_i, \gamma$: 가중치, k : 전체 사용자수

• m,n : 텍스트, 비텍스트 요소의 개수

• $f_i(x)$: 문서 X에 대한 텍스트 요소

• $g_i(x)$: 문서 X에 대한 비텍스트 요소

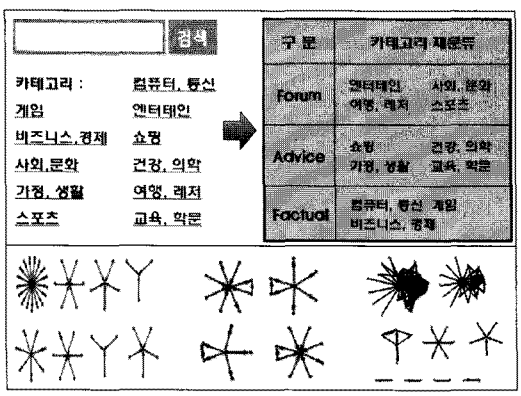
• indegree : 사용자 i가 다른 사용자 j로부터 받는 답변 채택수

• outdegree : 사용자 i로부터 다른 사용자 j에게 가는 질문수

텍스트 요소와 비텍스트 요소는 답변자에 의해 작성되어 지고 정해지는 값이다. 답변자의 능력에 따라 텍스트 요소와 비텍스트 요소가 답변의 품질에 얼마만큼 영향을 미칠 수 있느냐가 결정 된다 할 수 있다. 따라서 답변자의 영향력 또는 신뢰성이라 할 수 있는 사용자 중앙성 요소가 텍스트, 비텍스트 요소 모두에 영향을 미치기 때문에 곱 연산을 수행하여 Quality value를 도출하였다.

4. 실험 및 평가

본 장에서는 제안한 QualityRank 알고리즘의 실험 및 평가를 위해 네이버 지식iN 서비스에서 질의답변을 활용하였다. 네이버 지식IN에서 카테고리를 그림 3과 같이 Factual, Advice, Forum 클러스터로 분류하고 질의/답변자간 소셜 네트워크를 형성하였다[10]. 그림 3은 네이버 지식iN의 소셜 네트워크 구조이다.



(a) Factual (b) Advice (c) Forum

그림 3 네이버 지식iN의 소셜 네트워크 구조

4.1 데이터 셋

본 연구에서는 2010년 7월 한 달 간 네이버 지식IN에 입력된 질의답변을 수집하였다. 질의는 하루 동안 입력된 질의답변, 전체 네이버 지식IN에서 무작위로 선정된 질의답변, 질의에 대한 답변들 중 질문자가 "Best Answer"로 선택한 답변으로 이루어져 있다.

실험을 위해 학습용 데이터와 평가용 데이터로 구분하였다. 학습용 데이터는 전체 네이버 지식 IN에서 선택된 질문에 대한 답변들을 학습 집합으로 사용하고, 평가용 데이터는 질의의 특성상 Advice 클러스터의 답변이 객관적인 사실과 주관적인 의견으로 고루 분포되어 있어 클러스터내 질의에 적합하다고 판정된 답변을 사용했다. 각 답변에 대하여 신뢰도는 3점 척도(높음, 보통, 낮음)를 기준으로 문헌정보학 전공자들에 의해 평가되었다[1].

표본 추출 방식에 있어서는 전체 표본 프레임 구성에서 표본 선정의 용이성을 위해 체계적 표본 추출[1] 방식으로 데이터를 수집 하였다. 체계적 표집은 뽑은 사람의 주관이 배제된 상태에서 동등한 확률로 뽑히도록 표본을 추출하므로 객관적이며 체계적인 방법이다.

Advice 카테고리내 데이터 중 표본 추출한 질문/답변은 총 2,783개이다.

4.2 신뢰도 평가기준

질문의 유형에 따라 요구되는 답변의 수준이 다르기 때문에 먼저 질문에 따라 지식형 질문(전문적인 지식이 필요로 하는 범주)과 생활형 질문(생활 상식이나 신변잡기적인 내용을 묻는 범주)으로 구분한 후 질문 유형에 따라 신뢰도 평가 기준은 표 5와 같으며, 이 기준의 평가 항목 중 하나 이상을 만족시키면 기준에 부합되는 것으로 평가되었다.

본 연구에서는 전체 지식IN에서 체계적인 표본 추출 방식으로 선택된 질문에 대한 답변들을 학습 집합으로 사용하고, Advice 클러스터내 질의에 적합하다고 판정된 답변의 집합을 평가집합으로 사용했다.

표 6은 학습용 집합과 평가용 집합의 신뢰도 분포를 나타낸다.

4.3 가중치 산출

선정된 텍스트/비텍스트 요소, 소셜 네트워크 요소와 학습용 집합 데이터의 신뢰도 점수(높음=2, 보통=1, 낮음=0)를 통해 Pearson 상관계수로 표 7, 8, 9와 같이 가중치를 산출하였다. Pearson 상관계수는 식 (8)과 같다.

1) 체계적 표집(systematic sampling) : 확률적 표집 방식의 하나로, 전체 모집단의 크기를 N이라 하고 일정한 질서에 따라서 n크기의 표본을 추출하는 방식.

표 5 답변의 신뢰도 평가 기준

신뢰도 점수	지식형	생활형
높 음	- 공신력 있는 정확한 출처 - 객관적으로 확실한 근거 (이론/학문적 예시 등) - 논리적 설명 - 자료 첨부 (표, 그림, 사진) - 질문에 대한 핵심 답변	- 논리적인 개인의견 - 속담, 격언, 생활지식 등 학문적 근거는 없지만 상식적 답변 - 질문에 대한 핵심 답변
보 통	- 답변을 하였으나 근거부족	- 정확한 출처가 나오지는 않고 답변의 의견에 의존하나 어느 정도 논리적임
낮 음	- 비방, 욕설, 음란한 글 - 명예 훼손성 글 - 추측성 답변 - 근거가 없는 개인의견 - 질문과는 전혀 관련없는 답변 - 광고성 글	

표 6 데이터의 신뢰도 분포

구분	높음	보통	낮음
학습용	524	315	167
평가용	969	1065	749
전 체	1,493	1,380	916

$$r=\frac{\Sigma XY-\frac{\Sigma X\Sigma Y}{N}}{\sqrt{\left(\Sigma X^2-\frac{(\Sigma X)^2}{N}\right)\left(\Sigma Y^2-\frac{(\Sigma Y)^2}{N}\right)}} \quad (8)$$

• X : 텍스트/비텍스트, 소셜 네트워크 요소

• Y : 신뢰도 점수, N : 답변의 개수

가중치의 값이 양수이면 품질 요소의 값이 커질수록 신뢰도 점수도 높아진다는 것을 나타내고, 음수이면 품질 요소의 값이 작아질수록 신뢰도 점수가 높아진다는 것을 나타낸다. 상관계수가 0에 가까울수록 실제 신뢰도 점수와의 관련성이 낮음을 나타낸다.

표 7, 8, 9의 측정 결과를 보면 텍스트 요소에서 답변의 길이 요소(α_5)와 비텍스트 요소에서는 추천수(β_1)가 가장 높다. 소셜 네트워크 요소는 이 두 요소 보다 더 높은 상관계수가 산출된 것을 알 수 있다.

표 7 텍스트 요소의 가중치

구분	α_1	α_2	α_3	α_4	α_5
가중치	0.1204	0.0156	0.1058	-0.0209	0.2110

표 8 비텍스트 요소의 가중치

구분	β_1	β_2	β_3	β_4	β_5
가중치	0.2034	0.1371	-0.0251	0.1627	0.1889

표 9 소셜 네트워크 요소의 가중치

구분	γ
가중치	0.2437

4.4 알고리즘 평가

본 논문에서는 QualityRank 알고리즘을 통해 랭킹화된 질의답변 문서를 3점 척도 법으로 점수를 부여한 후 기존 정보검색 시스템 평가에 사용되는 Precision, Recall[11]과 NDCG[12]를 통해 알고리즘의 정확도를 비교 평가한다.

• 일반화된 정확률과 재현율

질의답변 문서에 대해 랭킹 성능을 평가하기 위한 정보검색 시스템 평가 척도로, 높은 등급의 문서가 상대적으로 낮은 등급의 문서보다 상위에 출현할 경우 높은 점수를 부여하여 성능을 측정하는 방식이다. 일반화된 정확률(P)와 재현율(R)의 정의는 다음과 같다.

$$P=\sum_{d\in R}\frac{r(d)}{N} \qquad R=\frac{\sum_{d\in R} r(d)}{\sum_{d\in D} r(d)}$$

• r(d) : 문서 d의 점수,

• R : D={d1,d2,....,dn}에서 검색된 문서집합(전체 답변)

• N : 검색된 문서 집합의 크기(전체 답변의 개수)

본 연구에서 다루는 답변은 3점 척도로 점수가 부여되었다. 정확율(P)을 이용하여 평균 정확율(AP)은 다음과 같이 계산할 수 있다.

$$AP=\frac{\sum_{r=1}^{N}(P(r)\times isrel(r))}{\text{적합한문서의개수}}$$

• P(r): 정확율, isrel(r) : 신뢰도 점수

• NDCG(Normalized Discounted Cumulative Gain)

질의답변 문서에 대해 랭킹 성능을 평가하기 위한 척도로, 높은 등급의 문서가 상대적으로 낮은 등급의 문서보다 상위에 출현할 경우 높은 점수를 부여하여 성능을 측정하는 방식이다.

$$NDCG_P=\frac{DCG_P}{IDCG_P}$$

• NDCG : Normalized Discounted Cumulative Gain

• IDCG : Ideal Discounted Cumulative Gain

IDCG는 가장 이상적인 랭킹점수를 모두 합한 값이며 DCG(Discounted Cumulative Gain)는 n개의 랭킹 결과물의 등급을 모두 합한 값이다. 추가하여 부여된 점수에는 비례하지만 낮은 랭킹의 결과에 대해서 점진적으로 패널티를 주기 위해 LOG 함수를 사용한다.

$$DCG_P = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2 i}$$

• P : 랭킹위치, rel : 점수

4.5 실험결과 및 분석

제안하는 알고리즘의 상대적인 문서 품질을 알아보기 위해 각 요소에 대한 알고리즘의 성능을 평균정확율과 재현율로 측정하였다. QualityRank 알고리즘과 기존 텍스트, 비텍스트 요소의 성능을 비교실험한 결과는 표 10과 같다.

정확률과 재현율로 실험한 결과는 그림 4와 같다.

그림 5는 수집된 데이터를 QualityRank 알고리즘을 통해 랭킹화를 거친 후 상위 10%에 대해 NDCG를 이용하여 알고리즘 성능을 비교하였다.

위 실험결과에서 지식문서 품질평가에 있어 본 논문에서 제안하는 QualityRank 알고리즘이 Textual과 Non-textual에 비해 평균 정확률에서 최대 7%, 랭킹 상위 10%에 대한 NDCG 측정결과에서는 최대 4%정도 개선된 것을 알 수 있다. 주목할 점은 기존의 Textual 속성만으로 평가했을 때와 Non-Textual한 속성만을 고려했을 때보다 두 개의 속성을 동시에 고려하여 문서의 품질을 평가했을 때가 더 높은 정확율과 재현율을 보였다.

정확율과 재현율이 높다는 것은 제안한 알고리즘이 지식검색 서비스 내에서 품질이 높은 문서를 찾는데 성능이 우수하다는 것을 의미한다. 결과적으로 지식 공유의 근본적 목적에 부합되도록 어떠한 형태든 양질의 질문/답변 문서를 평가하는 품질을 객관적 수치로 선별

표 10 평균정확률로 측정한 알고리즘 성능비교

구분	QualityRank	비텍스트	텍스트
평균 정확률	76.60	69.96	71.95

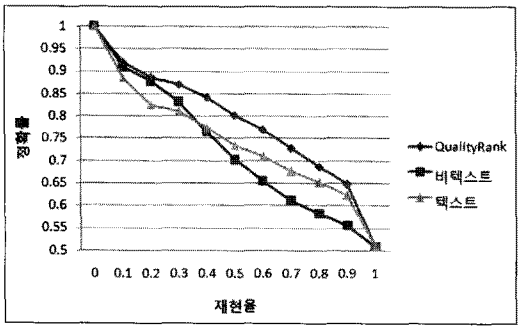
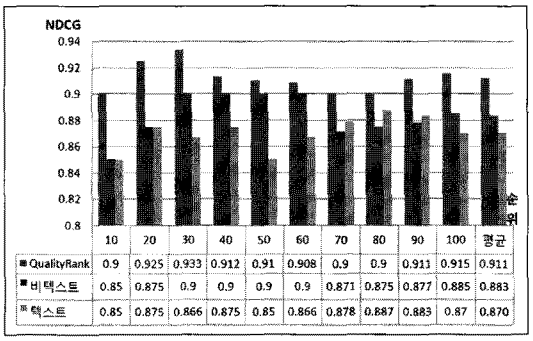


그림 4 QualityRank 알고리즘 성능 비교



	10	20	30	40	50	60	70	80	90	100	평균
QualityRank	0.9	0.925	0.933	0.912	0.91	0.908	0.9	0.9	0.911	0.915	0.911
비텍스트	0.85	0.875	0.9	0.9	0.9	0.9	0.871	0.875	0.877	0.885	0.883
텍스트	0.85	0.875	0.866	0.875	0.85	0.866	0.878	0.887	0.883	0.87	0.870

그림 5 NDCG를 이용한 QualityRank 성능 비교

가능함을 보여주는 것이라 할 수 있다. 선정된 품질지수를 통해 사용자가 원하는 지식 정보의 획득이 보다 용이해질 수 있을 것으로 판단된다. 또한, 향후연구에서는 보다 더 효과적인 요소 추출 방법을 모색하여 성능의 개선폭을 개선시킬 계획이다.

5. 결 론

본 논문에서는 비텍스트 정보와 텍스트 정보가 갖는 문제점을 완화시킬 수 있도록 텍스트 정보와 비텍스트 정보, 사용자 중앙성 정보를 동시에 고려하여 평가 알고리즘을 제안했다. 또한 네이버 지식 질의/답변 서비스에서 수집한 실제 문서들을 대상으로 하여 확률 기반의 제안한 문서 품질 알고리즘을 적용하였다. 일반화된 정확율과 재현율을 통해 성능을 평가한 결과 기존 텍스트 요소와 비텍스트 요소만을 각각 고려한 경우 보다 QualityRank 알고리즘을 사용했을 때 높은 결과를 나타냈다. 이는 질의/답변 문서를 평가하는 품질을 객관적 수치로 선별 가능함을 보여주는 것이라 할 수 있다.

또한, 품질의 지표로 지식검색 서비스의 만족도를 한층 더 향상시킬 수 있는 상당히 의미있는 성능 향상으로 판단된다.

향후연구로서 텍스트/비텍스트 특성 기반 QualityRank 알고리즘 성능을 개선하기 위한 방안이 필요하다. 보다 더 효과적인 요소 추출 방법을 모색하여 기존 방법 대비 알고리즘, 가중치 계산 등에 있어 성능의 개선폭을 향상시키는데 연구할 예정이다. 또한, 제안한 알고리즘은 지식검색 문서에 종속적인 것이 아니기 때문에, 지식검색 서비스에서의 문서 품질 평가뿐만 아니라 블로그, 제품 리뷰 등 다른 종류의 사용자 제작 문서의 품질평가에도 유용할 것이라 생각 된다. 일반적인 사용자 제작문서의 성능을 평가하고 효과적인 향상방안에 대해 연구가 필요할 것이다.

참 고 문 헌

[1] 박소연, 이준호, 전지운, "지식검색 서비스 개선을 위

한 문서의 적합도 및 신뢰도 분석", *한국문헌정보학회지*, vol.40, no.2, p.300, 2006년.
[2] 이정태, 송영인, 임해창, "신뢰도 자질을 이용한 지식 검색 문서의 품질 평가", *한국정보과학회 학술발표 논문집*, pp.63-65, 2007년 10월.
[3] Jiwoon Jeon, W.Bruce Croft, Joon Ho Lee, Soyeon Park, "A Framework to Predict the Quality of Answers with Non-Textual Features," *In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.228-235, 2006.
[4] 김희연, "정보사회에서의 지식과 지식검색에 대한 고찰", *정보통신정책*, vol.18, no.14, 통권398호, pp.6-8, 2006년 8월.
[5] Szuba T, "Computational Collective Intelligence," Wiley and Sons NY, 2001.
[6] M. V. Vieira, B. M. Fonseca, R. Damazio, P. B. Golgher, D. de Castro Reis and B. Ribeiro-Neto, "Efficient Search Ranking in Social Networks," *In Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM)*, pp.563-572, 2007.
[7] Freeman L, "Centrality in Social Networks : A Conceptual Classification," *Social Networks*, no.1, 1979.
[8] 김용학, "사회연결망 분석", 박영사, pp.7-36, pp.82-122, 2003년.
[9] Berger, A. L., Pietra, V.J.D., and Pietra S.A.D., "A maximum entropy approach to natural language processing," *Compt. Linguist.*m vol.22, no.1, pp.39-71, 1996.
[10] Lada A.Adamic, Jun Zhang, Eytan Bakshy, Mark S.Ackerman, "Knowledge Sharing and Yahoo Answers : Everyone Knows Something," *WWW 2008*, pp.667-670, April 2008.
[11] Kekalainen, J. and Jarvelin, K., "Using graded relevance assessments in IR evaluation," *Journal of the American Society for Information Science and Technology*, vol.53, no.13, pp.1120-1129, 2002.
[12] Kalervo Jarvelin, Jaana Kekalainen, "Cumulated Gain-based Evaluation of IR Techniques," ACM, pp.2-22, 2002.

김 덕 주
2000년 공군사관학교 전자과 학사. 2009년 국방대학교 전산정보학과 석사과정. 현재 국방대학교 전산정보학과 석사과정. 관심분야는 정보검색, 소셜 네트워크

박 건 우
1997년 충남대학교 컴퓨터과학과 학사. 2007년 연세대학교 컴퓨터과학과 석사. 2008년 국방대학교 전산정보학과 박사과정. 현재 국방대학교 전산정보학과 박사과정. 관심분야는 정보검색, 소셜 네트워크, 네트워크 보안

이 상 훈
1978년 성균관대학교 정보통신공학과 학사. 1989년 연세대학교 산업대학원 전산학과 석사. 1997년 일본 교토대학교 정보공학 박사. 서일대학 겸임교수, 충남산업대학교 교수, 일본 교토대학교 교환교수. 현재 국방대학교 전산정보학과 교수
관심분야는 정보검색, 데이터베이스, 미디어 융합