

# 신뢰도 자질을 이용한 지식검색 문서의 품질 평가

이정태 송영인 임해창  
고려대학교 컴퓨터학과  
{jtle, song, rim}@nlp.korea.ac.kr

## Quality Prediction of Knowledge Search Documents Using Text-Confidence Features

Jung-Tae Lee Young-In Song Hae-Chang Rim  
Dept. of Computer Science and Engineering, Korea University

### 요 약

불특정 사용자의 참여에 의해 정보가 생성되는 지식검색 서비스에서는 문서의 품질이 검색 만족도에 중요한 요소 중 하나이다. 지식검색 문서의 품질 평가에 관한 기존 연구는 조회 수나 추천 수 등의 비텍스트 정보를 이용하여 문서의 품질을 평가하고, 이를 검색 모형에 반영하여 검색 성능을 높이는데 집중하였다. 이러한 비텍스트 정보는 그 유용성이 실험을 통해 증명되었지만, 새로 작성된 문서와 같은 경우 심각한 자료 부족 문제가 발생할 수 있다는 단점이 있다. 본 논문에서는 이러한 비텍스트 정보의 자료 부족 문제를 완화할 수 있는 새로운 문서 품질 평가 자질로서 문서 내용의 신뢰성을 반영하는 신뢰도 자질을 제안한다. 제안하는 자질은 문서의 내용으로부터 직접 추출되며, 따라서 추천 수나 조회 수 등 서비스 사용자의 참여나 이용을 필요로 하는 비텍스트 자질보다 자료 부족 문제에 견고하다는 장점이 있다. 또한 제안하는 신뢰도 자질은 문서 품질 평가에 유용하다고 알려진 비텍스트 자질과 유사하거나 향상된 성능을 실험에서 보였으며, 추후 자질 추출 방법을 개선한다면 효과적인 품질 평가 자질로서 기능을 할 수 있을 것으로 기대된다.

### 1. 서론

지식검색 서비스란 사용자가 자발적으로 참여하여 쌍방향으로 질문과 답변을 하는 커뮤니티 기반의 서비스이다. 일반 검색엔진으로 찾기 힘든 정보를 원할 때 이를 질문으로 작성하면 다른 불특정다수의 사용자들로부터 답변을 얻을 수 있다는 강점 때문에 실제로 오늘날 네이버, 야후, 애플스 등 여러 포털의 대표적인 서비스로 정착하였다. 그러나 수많은 사용자들이 참여하게 되면서 그 어느 사용자라도 웹 브라우저만 있으면 답변을 작성할 수 있기 때문에 지식검색 문서의 품질을 해결하는 것이 시급한 상황이다. 이를 해결하기 위하여 지식검색 문서의 품질 평가에 관한 연구가 이루어지고 있지만 아직 미흡한 편이다.

지식검색 문서의 품질 평가에 관한 기존 연구에서는 문서의 추천 수나 조회 수 등 지식검색 서비스 제공자가 측정하는 비텍스트 정보를 이용하여 문서의 품질을 확률적으로 평가하는 방법을 제안하였고, 이를 검색 기법에 적용하여 검색성능을 향상시켰다[1]. 추천 수 또는 클릭 수 같은 비텍스트 정보는 비록 문서에 대한 간접적인 정

보라고 해도 문서의 품질을 평가하기 위한 자질로 유용하나, 다음과 같은 자료 부족 문제가 발생할 수 있다. 예를 들어 비텍스트 정보만 사용하여 문서의 품질을 평가할 경우, 상대적으로 뒤늦게 작성된 답변은 아무리 내용이 훌륭하다고 할지라도 이전에 작성된 답변에 비해 비텍스트 정보가 부족할 수밖에 없기 때문에 품질을 공평하게 평가할 수 없다는 단점이 있다.

본 논문에서는 기존 연구와 다르게 지식검색 문서의 품질을 내용의 신뢰도 측면에서 평가하는 방법을 제안한다. 이 방법은 높은 품질의 답변은 그 답변을 읽게 되는 사용자로 하여금 충분히 신뢰할 수 있을 정도로 최대한 성실하고 자세하게 작성되었고, 이에 비해 낮은 품질의 답변은 사용자가 신뢰할 수 없게끔 내용 면에서 불성실하고 불충분하게 작성되었을 것이라는 가정을 기반으로 한다.

본 연구에서는 문서 내용의 신뢰도를 측정하기 위한 새로운 자질을 제안한다. 또한 제안하는 신뢰도 자질의 유용성을 입증하기 위해 기존 비텍스트 정보를 다룬 연구[1]에서 진행한 방법과 동일하게 확률 기반의 문서 품질 평가 모델에 제안하는 자질을 적용하고 네이버 지식

검색 서비스에 실제로 존재하는 답변 문서를 대상으로 성능을 측정한다.

이후, 본 논문의 구성은 다음과 같다. 2장에서는 기존 연구에 대하여 논하고, 3장에서는 본 논문에서 제안하는 지식검색 문서의 품질 평가를 위한 신뢰도 자질에 대해 상세 기술한다. 4장에서 본 연구에서 사용하는 문서의 품질 평가를 위한 모델에 대해 설명하고, 5장에서는 실험 데이터 및 실험 결과에 관하여 기술한다. 마지막으로 결론 및 향후연구에 대하여 언급한다.

## 2. 기존 연구

본 논문에서 다루고 있는 문서의 품질 평가에 관한 연구는 학생들이 작성한 논술을 컴퓨터를 이용하여 자동으로 채점하는 논술 자동채점(Automated Essay Grading)에 관한 연구와 관련이 있다. 논술 자동채점 프로그램의 가장 대표적인 예로 Program Essay Grade(PEG)[2]를 들 수 있다. PEG는 사람이 논술을 채점할 때와 같이, 주어진 논술이 얼마나 잘 작성되었는지에 초점을 맞춘다. 이를 기반으로, 컴퓨터가 쉽게 측정할 수 있는 자질을 사용한다. 예를 들어 논술에 쓰인 문체의 유창함은 쉽게 측정하기 힘들지만, 논술의 길이를 이용하여 이를 추정할 수 있다. PEG의 접근방법은 단순하지만 예상보다 훨씬 높은 성능을 보인 것으로 알려져 있다. PEG에서 사용된 자질 중에서는 논술의 길이가 매우 중요하다고 보고되었다.

인터넷이 발전하면서 웹페이지 또한 품질 평가 연구의 중요한 대상이 되어왔다. 이에 관련된 연구 중 상당한 양이 웹페이지의 링크 즉, 웹의 구조적 특성을 분석하는 방식이었다. 대표적인 예로 PageRank[3] 또는 HITS[4]와 같은 링크 기반 방식이 제안되었다. 이러한 방식은 주어진 한 웹페이지를 향한 링크의 개수가 그 웹페이지의 평판(popularity)을 가리키는 척도라는 것에 기반을 둔다. 하지만 이러한 방식은 다음과 같은 자료 부족 문제를 안고 있다. 새로 생성된 웹페이지의 경우 검색결과 상위에 나타나는 웹페이지에 비해 가리키는 링크가 상대적으로 적을 수밖에 없다. 이를 해결하기 위해 링크 이외에 웹페이지 내용으로부터 정보 대 잡음 비율(information-to-noise ratio) 등을 계산하여 문서의 품질을 평가하는 방식[5,6] 또한 제안되었다. 그러나 아직 웹페이지 품질에 대한 표준적인 평가기준은 설정되지 않은 실정이다.

최근에는 사용자가 자발적으로 문서 구축에 참여하는 지식검색 서비스가 빠른 속도로 여러 국내 포털의 대표적인 서비스로 자리를 잡았지만, 지식검색 서비스의 질적 향상을 위한 연구나 지식검색 문서의 평가에 대한 국내 연구는 미흡한 실정이다. 국내 관련 연구 중에서는 지식검색 문서의 평가 기준을 제시한 연구[7]를 들 수 있다. 구체적으로 이 연구에서는 지식검색 문서를 구성하는 질문과 답변의 적합도 평가 기준과 답변의 신뢰도 평가 기준을 제시하고, 또한 이러한 평가 기준에 근거하여 지식검색 문서의 실제 평가를 수행하였다.

지식검색 문서의 품질 평가를 위한 연구로는 서론에서 언급한 추천 수 또는 조회 수 등 서비스 제공자 측에서 수집하는 비텍스트 정보를 이용하여 지식검색 문서의 품질을 평가하는 방법을 제안한 연구[1]를 들 수 있다. 이 방법의 경우 추천 수나 조회 수 등 지식검색 문서의 답변과 관련된 총 13가지의 비텍스트 정보를 활용하여 기계학습 기반의 확률적 품질 평가 분류기를 만들어 품질 지수를 측정하고, 이를 검색 모델에 반영하여 검색 성능을 향상시켰다. 이러한 비텍스트 정보는 품질 평가에 충분한 양을 수집하는데 있어 시간을 요한다는 점이 단점으로 작용할 수 있다. 예를 들어 상대적으로 나중에 작성된 답변은 아무리 내용의 품질이 높다고 한들 이전에 작성된 답변에 비해 비텍스트 정보의 양이 부족할 수밖에 없고, 따라서 공평한 평가가 어려운 단점이 있다.

문서 품질 평가에 관한 연구는 의견 요약(Opinion Summarization) 분야에서도 이루어지고 있다. 의견 요약 분야의 최근 연구동향은 쇼핑몰 사용자들이 어떤 특정 제품이나 서비스에 대하여 직접 작성한 리뷰로부터 제품의 기능요소에 관한 의견을 요약하는 방법에 관한 연구가 활발하게 이루어지고 있다. 그중 문서 품질을 고려하는 연구로는 리뷰의 품질을 기계학습으로 평가하여 낮은 품질의 리뷰를 판별 및 여과하고, 이를 요약 기법에 반영하는 방법을 제안한 연구[8]를 들 수 있다. 이 연구에서는 리뷰 문서의 품질을 평가하기 위해서 리뷰 내용에 사용된 단어의 총 개수, 문장의 평균 길이, 주관적 문장(subjective sentence)<sup>1</sup>의 비율 등을 자질로 사용하였다.

## 3. 지식검색 문서의 품질 평가를 위한 신뢰도 자질

본 논문에서는 지식검색 문서의 품질 평가를 위해 신뢰도 자질이라고 정의한 새로운 자질을 제안한다.

지식검색 서비스를 이용하는 그 어느 사용자도 웹 브라우저만 있다면 다른 사용자가 올린 질문에 답변을 할 수 있기 때문에 지식문서의 품질은 천차만별이다. 문서의 품질을 평가하기 위한 기준은 물론 주관적인 것이지만, 본 연구에서는 주어진 질문에 대하여 사용자가 읽고 충분히 신뢰할 만큼 성실하고 자세히 기술된 답변을 높은 품질의 답변이라고 보고, 이에 비해 신뢰하기 어려운 정도로 불성실하게 부실하게 작성한 답변을 낮은 품질의 답변이라고 본다.

문서의 내용이 얼마나 성실하고 꼼꼼하게 작성되었는지 따라서 그 문서를 읽는 사용자로 하여금 내용에 얼마나 신뢰할 수 있는지를 측정하는 매우 어렵다. 관점에 따라 복잡하고 시간을 요하는 자연어처리 기술을 이용하여 문서를 분석하는 과정이 필요하다고 볼 수도 있다. 하지만 본 연구에서는 지식검색 서비스가 웹에서 제공되는 서비스라는 점에 착안하여 컴퓨터가 계산하기 쉬운 변수들을 자질로 사용한다. 자질들 중 대부분은 어휘사전(lexicon) 기반으로 측정하게 된다. 즉, 높은 품질의 답

<sup>1</sup> 의견, 감정 등을 기술한 문장

변에서 주로 출현하는 단어와 낮은 품질의 답변에서 자주 출현하는 단어로 미리 사전을 구축하고, 사전에 기록된 단어가 답변에 출현한 비율을 측정하여 자질로 활용한다.

본 논문에서 제안하는 신뢰도 자질의 구성은 다음과 같다.

- **연결어의 출현 비율** : 문장의 연결을 위해 사용되는 연결어가 주어진 답변에서 차지하는 비율을 의미한다. 성실한 문서의 경우, 근거를 논리 정연하게 설명하기 위해서 연결어를 빈번하게 사용할 것이고 이는 신뢰도를 향상시킬 것이라는 가정에 의해 자질로 선정되었다.
- **추정어의 출현 비율** : 추정의 느낌이 강한 문장에서 주로 사용되는 어휘가 주어진 답변에서 차지하는 비율을 의미한다. 추정어가 많이 사용되었을 경우 뚜렷한 확신 없이 글을 작성하였을 가능성이 있다고 이는 곧 신뢰도를 저하시킨다는 가정 하에 자질로 사용되었다.
- **개인의견 단어의 출현 비율** : 개인적 의견이 강한 문장에서 주로 나타나는 단어가 주어진 답변에서 차지하는 비율을 의미한다. 개인의견 단어가 빈번하게 사용될수록 문서가 객관적인 사실보다 주관적인 의견의 표현을 주로 담고 있을 가능성이 있기 때문에 신뢰도를 저하시킨다는 가정에 의해 자질로 선정되었다.
- **광고성 단어의 출현 비율** : 광고 메일이나 문자메시지 등에 주로 쓰이는 단어가 주어진 답변에서 차지하는 비율을 의미한다. 광고성 어휘가 빈번하게 사용되면 질문에 관련된 답변이라 할지라도 불성실한 답변 또는 질문에 적합하지 않은 답변일 가능성이 크기에 신뢰도를 저하시킨다고 가정한다.
- **가치판단어의 출현 비율** : 특정 객체에 대한 가치를 판단하기 위해 사용되는 가치판단어가 주어진 답변에서 차지하는 비율을 말한다. 성실하게 자신의 의견을 강조하거나, 어떤 주제에 대한 내용을 평가하고 근거를 제시하기 위해 자주 사용될 것이고, 따라서 신뢰도를 향상시킬 것이라는 가정 하에 자질로 사용하였다.
- **이모티콘의 출현 비율** : 주로 감정을 나타내는 이모티콘이 주어진 답변에서 차지하는 비율을 의미한다. 이모티콘이 빈번하게 나타난 답변일수록 가독성을 저하시키고, 이어 답변의 신뢰도를 저하시킬 것이라는 가정 하에 사용되었다.
- **웹사이트 주소 출현 여부** : 참조한 웹사이트의 주소가 답변 내에 출현했는지의 여부를 의미한다. 주소가 있을 경우 객관적으로 정보의 출처를 밝히는 답변일 가능성이 높고, 따라서 신뢰도가 향상될 것이라는 가정에 기반을 두어 사용하였다.
- **답변의 길이** : 답변이 무조건 길다고 하여 성실하다는 보장은 없지만, 본 연구에서는 주어진 답변의

길이가 길수록 질문에 대하여 성실하고 최대한 자세하게 기술한 답변이라고 가정한다.<sup>2</sup>

연결어나 추정어 등의 사전은 높은 품질의 답변에서 주로 나오는 단어 목록과 낮은 품질의 답변에서 가장 빈번하게 출현하는 단어 목록에서 적당한 단어를 학생 두 명이 직접 사전을 구축하였다. 사전에 기록될 단어를 뽑을 때에는 두 명의 동의하에서만 뽑았다. 그리고 이모티콘의 경우, 광고성 문자메시지에 주로 출현하는 어구 위주로 사전을 구축하였다. 각 어휘사전 기반 자질의 예는 <표 1>에 나타나 있다.

<표 1> 사전기반 자질의 예제

자질	예제
연결어	그러므로, 왜냐하면, 그러나
추정어	글쎄요, 아마도, 잘 모르겠네요
개인의견 단어	제 생각엔, 제가 듣기로는
광고성 단어	성인, 무료, 대출
가치판단어	빠른, 매우, 우수, 저렴, 더욱
이모티콘	^^, ^^, ^^, ^^

#### 4. 품질 평가 모델

본 연구에서는 비텍스트 정보를 다룬 연구[1]에서 사용한 것과 동일한 분류 기반의 확률적 문서 품질 평가 모델을 구현하고, 이에 신뢰도 자질을 적용하여 자질의 유용성을 보이고자 한다. 이 모델은 최대 엔트로피(maximum entropy) 모델[9]에 기반을 둔다.<sup>3</sup>

주어진 답변을 문서  $x$ 라고 하고 이  $x$ 라는 문서에 매길 수 있는 품질 등급을  $y = \{good, medium, bad\}$ 라고 하자. 본 품질 평가 모델의 목적은 조건부 확률  $p(y = good|x)$  즉, 주어진 문서가 높은 품질의 문서일 확률을 구하는 것이다. 최대 엔트로피 모델을 이용하면  $p(y|x)$ 는 최대 엔트로피 모델을 이용하여 다음과 같이 계산될 수 있다.

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

위 수식에서  $f_i(x, y)$ 는  $i$ 번째 자질이고,  $\lambda_i$ 는  $i$ 번째 자질의 가중치<sup>4</sup>이며,  $Z$ 는 모든 가능한  $y$ 에 대하여  $p(y|x)$ 을 더했을 때의 값이 1이 되도록 하기 위한 정규화 인수이다.

<sup>2</sup> 길이 자질은 기존의 논술 자동채점 연구[2] 및 비텍스트 정보를 이용한 지식검색 문서의 품질 평가 연구[1] 등에서 유용한 자질임이 입증되었다.

<sup>3</sup> 본 연구에서는 Zhang Le의 최대 엔트로피 모델링 툴킷([http://homepages.inf.ed.ac.uk/s0450736/maxent\\_toolkit.htm](http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.htm))을 사용하여 품질 평가 모델을 구현하였다.

<sup>4</sup> 가중치를 구할 때에는 Limited-Memory Variable Metric 기법[10]을 사용한다.

<표 2> 답변의 신뢰도 평가 기준 [7]

신뢰도	지식형	생활형
높음	<ul style="list-style-type: none"> <li>• 공신력 있는 정확한 출처</li> <li>• 객관적으로 확실한 근거(이론적, 학문적인 예시 등)</li> <li>• 논리적 설명</li> <li>• 자료 첨가(표, 그림, 사진)</li> <li>• 질문에 대한 핵심 답변</li> </ul>	<ul style="list-style-type: none"> <li>• 논리적인 개인의견</li> <li>• 속담, 격언, 생활지식 등 학문적 근거는 없지만 상식적</li> <li>• 질문에 대한 핵심 답변</li> </ul>
보통	<ul style="list-style-type: none"> <li>• 답변을 하였으나 근거부족</li> </ul>	<ul style="list-style-type: none"> <li>• 정확한 출처가 나오지는 않고 답변의 의견에 의존 하나 어느 정도 논리적임</li> </ul>
낮음	<ul style="list-style-type: none"> <li>• 비방, 욕설, 음란한 글</li> <li>• 명예훼손성 글</li> <li>• 추측성 답변</li> <li>• 근거가 없는 개인의견</li> <li>• 질문과는 전혀 관련 없는 답변</li> </ul>	

최대 엔트로피 기반 모델은 다양한 정보를 자질로 결합할 수 있는 장점을 가지면서 조건부 확률을 추정할 수 있다. 본 품질평가 모델이 출력하는 확률은 추후 언어모형(language model) 기반 검색모델<sup>5</sup>에 적용할 수 있다. 실제로 비텍스트 정보, 문서의 내용 정보, 평판(popularity), 시간 정보 등으로 확률을 추정하고, 이를 언어모형 기반 검색모델에 적용하여 검색성능을 향상시킨 관련 연구들이 존재한다[1,6,12,13].

## 5. 실험

### 5.1. 실험 데이터

본 연구에서 사용한 실험 데이터 집합은 지식검색 문서의 평가 기준을 제시한 기존 연구[7]에서 구축된 것으로서, 비텍스트 정보를 이용한 연구[1]에서 사용한 것과도 동일한 집합이다. 본 데이터 집합은 네이버 지식검색 서비스<sup>6</sup>에서 수집된 문서로 이루어져 있으며, 각 문서는 사용자가 올린 질문과 이에 대해 타사용자가 작성한 답변으로 이루어져 있다.

각 답변에 대하여 신뢰도가 3점 척도(높음, 보통, 낮음)를 기준으로 문헌정보학 전공자들에 의해 평가되었다. 질문의 유형에 따라 요구되는 답변의 수준이 다르기 때문에, 먼저 지식검색 문서에 포함된 질문에 따라 지식형 질문(전문적인 지식을 필요로 하는 범주)과 생활형 질문(생활 상식이나 신변잡기적인 내용을 묻는 범주)으로 구분한 후 질문 유형에 따라 신뢰도 평가 항목이 차별화되어 적용되었다. 구체적인 답변의 신뢰도 평가 기준은 <표 2>와 같으며, 이 기준의 평가 항목 중 하나 이상을 만족시키면 기준에 부합되는 것으로 평가하였다.

<표 3> 및 <표 4>는 실험 데이터에 실제 존재하는 답변 중 신뢰도가 높음 또는 낮음으로 책정된 예를 하나씩

첨부한 것이다.

<표 3> 신뢰도가 높음으로 매겨진 답변의 예

질문 제목	바스티안같은 게임좀가르쳐주세요
질문 본문	바스티안이 짬있었는데 유료라 하지 못하고있네요...바스티안같은 그런 게임좀 소개 시켜주세요..^^
답변 제목	re: 바스티안같은 게임좀가르쳐주세요
답변 본문	바스티안을 하지 못하시구 계시는 이유가 '유료' 때문이라면 바스티안을 다시 시작해 보시는게 어떨까요 ^^? 바스티안은 2004년 4월 8일 무료화 되었답니다 홈페이지 가보시면 아시겠지만 만약 셉이 바티아가 아니시라면 바티아로 셉 이전 신청을 하셔야 할겁니다 ^^ (이하 생략)
신뢰도가 높은 이유	객관적으로 확실한 근거(날짜 제시), 질문에 대한 핵심 답변

<표 4> 신뢰도가 낮음으로 매겨진 답변의 예

질문 제목	조선일보랑 중앙일보중에서
질문 본문	조선일보랑 중앙일보중에 어떤거 보는데 낫까요? (이하 생략)
답변 제목	둘다 괜찮은데요..저는 조선일보를 본답니다.
답변 본문	혹자들은 뭐 매국노의 신문이니 어찌나 하지만 제가 보기에는 스타일이 가장 맘에 들거든요. 그리고 가장 많이 보는 신문이라...가장 대중적이랍니다.
신뢰도가 낮은 이유	근거가 없는 개인의견

실험을 위해 889개의 답변을 학습 집합으로, 1700개의 답변을 테스트 집합으로 사용하였다. 학습 집합과 테스트 집합의 구성은 <표 5>에 나타나 있다.

<sup>5</sup> 언어모형 기반 검색모델에 관한 자세한 내용은 [11]을 참조

<sup>6</sup> <http://kin.naver.com/>

<표 5> 실험 데이터의 구성

집합	상	중	하
학습	599 (67%)	212 (24%)	78 (9%)
테스트	1099 (65%)	393 (23%)	208 (12%)

### 5.2. 평가 척도

본 연구에서는 품질 평가 모델이 출력하는 조건부 확률 즉, 답변이 좋은(신뢰도가 높은) 문서일 확률로 순위가 매겨진 답변 목록을 대상으로 일반화된 재현율과 정확률(generalized recall and precision)[14]로 모델의 성능을 측정한다. 일반화된 재현율과 정확률은 다등급(multi-grade) 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 재현율  $gR$ 과 정확률  $gP$ 의 정의는 다음과 같다.

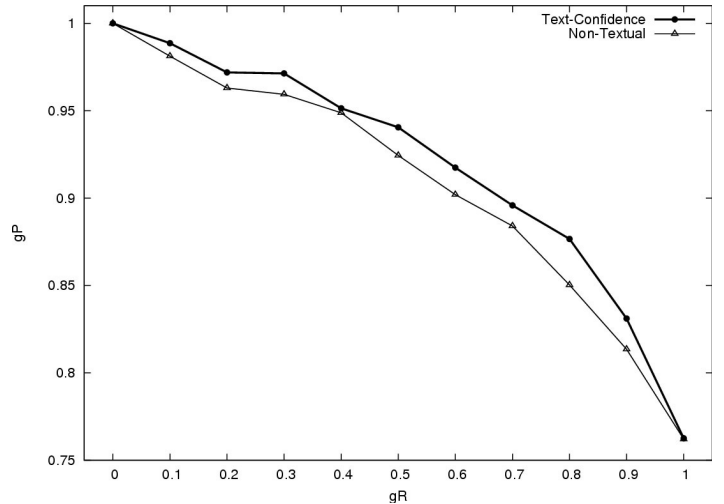
$$gP = \sum_{d \in R} \frac{r(d)}{N} \quad gR = \frac{\sum_{d \in R} r(d)}{\sum_{d \in D} r(d)}$$

여기에서  $r(d)$ 는 문서  $d$ 의 점수,  $R$ 은 집합  $D = \{d_1, d_2, \dots, d_n\}$ 에서 검색된 문서 집합(본 연구에서는 전체 답변), 그리고  $N$ 은 검색된 문서 집합의 크기(본 연구에서는 전체 답변의 개수)이다. 본 연구에서 다루는 답변은 3점 척도로 신뢰도가 부여되었기 때문에 전통적인 재현율과 정확률보다는 일반화된 재현율과 정확률이 본 실험 환경에 더 적합한 평가 척도라 할 수 있다.

### 5.3. 실험 결과 및 분석

앞서 기존 연구를 살펴본 바와 같이, 가장 최근의 지식검색 문서 품질 평가 방법은 비텍스트 정보를 이용한 방법[1]이다. 본 실험 환경의 실험 데이터와 품질 평가 모델은 모두 [1]에서 구축된 것과 동일하기 때문에 비텍스트 자질과 제안하는 신뢰도 자질 간의 성능 비교가 가능하다. 신뢰도 자질과 비텍스트 자질<sup>7</sup>의 성능 비교 실험 결과는 <그림 1> 및 <표 6>과 같다. (<그림 1>에서 'Text-Confidence'는 신뢰도 자질을, 'Non-textual'은 비텍스트 자질을 의미한다.)

7 비텍스트 자질로 사용한 정보는 답변 길이, 조회수, 답변수, 사용자 추천수, 사용자 비추천수, 출력회수, 스크랩회수, 답변자 채택률, 질문자 부여 평점, 답변자 내공, 답변자 전문가 지수, 관리자 추천 여부 및 스폰서 답변 여부 등 13가지로 [1]에서 사용한 것과 동일하다.



<그림 1>  $gR$ - $gP$  그래프로 측정된 성능

<표 6> 평균  $gP$ 로 측정된 성능

자질	평균 $gP$ (%)
신뢰도	91.77
비텍스트	90.72

위 실험 결과에서 지식검색 문서 품질 평가에 있어서 본 논문에서 제안하는 신뢰도 자질이 비텍스트 자질에 비해 성능을 개선하였음을 알 수 있으나, 그 개선의 폭은 약 1% 정도에 그쳐 큰 의미를 부여하긴 어렵다. 하지만 본 실험에서 사용한 신뢰도 자질의 추출이 수작업으로 구축된 소량의 어휘사전에 전적으로 의존한다는 한계점을 감안해볼 때, 그 유용성이 이미 입증된 비텍스트 자질과 동등하거나 나은 성능을 보여주었다는 것은 신뢰도 자질의 유용성을 입증하는 것이라 볼 수 있다. 신뢰도 자질 추출을 위하여 더 정교한 방법을 사용한다면, 문서 품질 평가에서 보다 높은 성능 향상을 이룰 수 있을 것이라 생각된다.

본 연구는 몇 가지의 한계를 지니고 있다. 먼저 실험 데이터의 부족으로 실험 결과의 신뢰성이 다소 떨어진다는 점이다. 신뢰도 자질을 추출하기 위해 수작업에 의한 어휘사전에 의존한다는 점 또한 이 연구의 한계이다. 이는 실험 데이터의 크기가 늘어나거나 또는 새로운 도메인의 실험 데이터의 문서 품질 평가가 요구될 때 추가적인 어휘사전 구축 작업을 필요로 할 가능성이 있다. 이러한 한계는 제안하는 방법의 가장 큰 단점으로 생각되며, 추후 연구에서 개선되어야 할 것으로 판단된다.

## 6. 결론

본 논문에서는 사용자의 참여에 의해 작성된 지식검색 문서의 품질을 평가하기 위한 자질로서 신뢰도 자질을 제안하였다. 또한 네이버 지식검색 서비스에서 수집한 실제 문서들을 대상으로 하여 확률 기반의 문서 품질 평가 모델에 제안한 자질을 적용하였고, 실험 결과를 통해

신뢰도 자질이 지식검색 문서의 품질 평가에 있어 유용함을 입증하였다. 기존 비텍스트 자질의 성능을 신뢰도 자질이 크게 개선하지는 못하였지만, 작은 어휘사전에 기반을 두어 비교적 단순한 방법으로 신뢰도 자질을 추출하여 사용하였음을 감안할 때, 상당히 의미 있는 성능 향상으로 판단된다.

또한, 본 연구에서 제안한 신뢰도 자질은 지식 검색 분야 문서에 종속적인 특징에 기반을 둔 것이 아니기 때문에, 지식검색 서비스에서의 문서 품질 평가뿐만 아니라 블로그, 제품 리뷰 등 다른 종류의 사용자 제작 문서의 품질 평가에도 유용할 것이라고 생각한다. 향후 연구로써, 신뢰도 자질 및 비텍스트 자질에 기반한 품질 평가 모델의 결과를 실제 검색 모델에 반영하여 일반적인 사용자 제작 문서의 검색 성능을 효과적으로 향상시킬 방안에 대해 연구하고자 한다.

### 감사의 글

본 연구는 NHN(주)의 지원을 받았습니다. 그러나 본 논문에서 제시된 의견이나 결론, 또는 권고 등은 온전히 저자들의 것이며, 반드시 지원회사의 입장을 대변하는 것은 아닙니다.

### 참고 문헌

- [1] J. Jeon, W. B. Croft, J. H. Lee, and S. 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.
- [2] E. B. Page, Computer Grading of Student Prose, Using Modern Concepts and Software, Journal of Experimental Education, Vol. 62, No. 2, pp. 127-142, 1994.
- [3] S. Brin and L. Page, The Anatomy of a Large-scale Hypertextual Web Search Engine, In Proceedings of the 7th International Conference on World Wide Web 7, pp. 107-117, 1998.
- [4] J. M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, J. ACM, Vol.46, No.5, pp. 604-632, 1999.
- [5] X. Zhu and S. Gauch, Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web, In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 288-295, 2000.
- [6] Y. Zhou and W. B. Croft, Document Quality Models for Web Ad Hoc Retrieval, In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 331-332, 2005.
- [7] 박소연, 이준호, 전지운, 지식 검색 서비스 개선을 위한 문서의 적합도 및 신뢰도 분석, 한국문헌정보학회지, 제40권, 제2호, 299-314쪽, 2006.
- [8] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, M. Zhou, Low-Quality Product Review Detection in Opinion Summarization, In Proceedings of 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 334-342, 2007.
- [9] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, A Maximum Entropy Approach to Natural Language Processing, Comput. Linguist. Vol.22, No.1, pp. 39-71, 1996.
- [10] R. Malouf, A Comparison of Algorithms for Maximum Entropy Parameter Estimation, In Proceedings of the Conference on Computational Natural Language Learning, pp. 49-55, 2002.
- [11] J. M. Ponte and W. B. Croft, A Language Modeling Approach to Information Retrieval, In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 275-281, 1998.
- [12] W. Kraaij, T. Westerveld, and D. Hiemstra, The Importance of Prior Probabilities for Entry Page Search, In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 27-34, 2002.
- [13] X. Li and W. B. Croft, Time-based Language Models, In Proceedings of the 12th ACM International Conference on Information and Knowledge Management, pp. 469-475, 2003.
- [14] J. Kekalainen and K. Jarvelin, 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.