텍스트 신뢰도 자질 기반 지식 질의응답 문서 품질 평가 모델

(Text-Confidence Feature Based Quality Evaluation Model for Knowledge Q&A Documents)

이 정 태 † 송 영 인 †† 박 소 영 ††† 임 해 창 ††††
(Jung-Tae Lee) (Young-In Song) (So-Young Park) (Hae-Chang Rim)

요 약 불특정 다수 사용자가 정보를 생성하는 지식 질의응답 서비스에서는 문서의 품질이 검색결과 만족도에 중요한 요소 중 하나이다. 지식 질의응답 문서의 품질 평가에 관한 기존 연구는 조회 수와 추천 수 등의 비텍스트 정보를 이용하여 문서의 품질을 평가하고, 이를 검색 모형에 반영하여 검색 성능을 높이는데 집중하였다. 이러한 비텍스트 정보는 그 유용성이 실험을 통해 증명되었다. 그러나 비텍스트 정보를 이용하여 새로 작성된 문서의 품질을 평가할 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 내용 자체에 대한 신뢰성을 반영하는 신뢰도 자질을 제안한다. 제안하는 자질은 문서의 내용으로부터 직접 추출되며, 따라서 추천 수나 조회 수 등 서비스 사용자의 참여를 간접적으로 필요로 하는 비텍스트 자질보다 자료 부족 문제에 견고하다는 장점이 있다. 또한 제안하는 신뢰도 자질은 문서 품질 평가에 유용하다고 알려진 비텍스트 자질과 유사하거나 향상된 성능을 실제 지식 질의응답 문서를 대상으로 한 실험에서 보였으며, 추후 효과적인 품질 평가 자질로서 지식 질의응답 서비스의 성능향상에 기여를 할 수 있을 것으로 기대된다.

키워드 : 지식 질의응답 서비스, 문서 품질 평가 모델, 신뢰도 자질

Abstract In Knowledge Q&A services where information is created by unspecified users, document quality is an important factor of user satisfaction with search results. Previous work on quality prediction of Knowledge Q&A documents evaluate the quality of documents by using non-textual information, such as click counts and recommendation counts, and focus on enhancing retrieval performance by incorporating the quality measure into retrieval model. Although the non-textual information used in previous work was proven to be useful by experiments, data sparseness problem may occur when predicting the quality of newly created documents with such information. To solve data sparseness problem of non-textual features, this paper proposes new features for document quality prediction, namely text-confidence features, which indicate how trustworthy the content of a document is. The proposed features, extracted directly from the document content, are stable against data sparseness problem, compared to non-textual features that indirectly require participation of service users in order to be collected. Experiments conducted on real world Knowledge Q&A documents suggests that text-confidence features show performance comparable to the non-textual features. We believe the proposed features can be utilized as effective features for document quality prediction and improve the performance of Knowledge Q&A services in the future.

Key words : Knowledge Q&A service, Document quality prediction model, Text-confidence features

· 본 연구는 NHN(주)와 2단계 BK21사업의 지원비를 받았음
· 이 논문은 제19회 한글 및 한국어 정보처리 학술대회에서 '신뢰도 자질을 이용한 지식검색 문서의 품질 평가'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 고려대학교 컴퓨터전파통신공학과
jtlee@nlp.korea.ac.kr
†† 정 회 원 : 고려대학교 컴퓨터학과
song@nlp.korea.ac.kr
††† 종신회원 : 상명대학교 디지털미디어학부 교수
ssoya@smu.ac.kr
†††† 종신회원 : 고려대학교 컴퓨터전파통신공학부 교수
rim@nlp.korea.ac.kr

논문접수 : 2008년 1월 3일
심사완료 : 2008년 8월 29일

정보과학회논문지: 소프트웨어 및 응용 제35권 제10호(2008.10)

1. 서 론

지식 질의응답 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, 엠파스 등 여러 포탈의 대표적인 서비스로 정착하였다. 그러나 수많은 사용자들의 참여로 인해 검증되지 않은 답변이나 출처가 불분명한 주관적인 답변이 범람함에 따라 지식 질의응답 문서(이하 지식문서)의 품질을 해결하는 것이 시급한 상황이다. 답변의 품질을 자동으로 측정하여 이를 기존의 지식문서 검색 기법에 반영한다면 저품질의 지식문서가 서비스 사용자에게 보일 가능성을 줄임과 동시에 고품질의 문서가 보일 가능성을 올릴 수 있으나, 아직 지식문서의 품질 평가에 관한 연구는 대체로 미흡한 편이다.

지식문서의 품질 평가에 관한 기존 연구[1]에서는 문서의 추천 수나 조회 수 등 지식 질의응답 서비스 제공자가 측정하는 비텍스트 정보를 이용하여 문서의 품질을 확률적으로 평가하는 방법을 제안하였고, 이를 검색 기법에 적용하여 검색성능을 향상시켰다. 추천 수 또는 클릭 수 등 기존 연구에서 사용된 비텍스트 정보는 비록 문서에 대한 간접적인 정보라고 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 지식문서의 품질을 내용의 신뢰도 측면에서 평가하는 방법을 제안한다. 이 방법은 고품질의 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 텍스트 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 질의응답 서비스에 실제로 존재하는 답변 문서를 대상으로 하여 성능을 측정한다.

이후, 본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 기존 연구에 대하여 논하고, 3장에서는 본 연구에서 사용한 지식문서 집합에 대해 설명한다. 4장에서는 본 논문에서 제안하는 지식문서의 품질 평가를 위한 신뢰도 자질 및 문서의 품질 평가 모델에 대해 설명한다. 5장에서는 실험 결과에 관하여 기술하고, 마지막으로 6장에서 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 문서 품질 평가 연구의 중요한 대상이 되어왔지만 아직 웹페이지의 품질에 대한 표준적인 평가기준은 설정되지 않은 실정이다. 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 검색결과의 상위에 나타나는 웹페이지에 비해 가리키는 링크가 상대적으로 적을 수밖에 없다.

이를 해결하기 위해 링크 이외에 웹페이지의 내용으로부터 측정된 정보 대 잡음 비율(information-to-noise ratio) 등의 텍스트 자질을 이용하여 문서의 품질을 평가하는 방식[5,6]이 제안되었다. 그러나 이러한 방법은 평가 데이터가 소량이거나 TREC 웹 문서 컬렉션에 특화된 방법이기 때문에 일반 사용자가 직접 생성하는 지식문서의 품질평가에 그대로 적용하기에는 힘들다.

최근에는 사용자가 자발적으로 문서 구축에 참여하는 지식 질의응답 서비스가 빠른 속도로 여러 포털 사이트의 대표적인 서비스로 자리를 잡았지만, 지식 질의응답

서비스의 질적 향상을 위한 연구나 지식문서의 평가에 대한 국내 연구는 미흡한 실정이다. 국내 관련 연구 중에서는 지식문서의 평가 기준을 제시한 연구[7]를 들 수 있다. 구체적으로 이 연구에서는 지식문서를 구성하는 질문과 답변의 적합도 평가 기준과 답변의 신뢰도 평가 기준을 제시하고, 또한 이러한 평가 기준에 근거하여 지식문서의 실제 평가를 수행하였다.

지식문서의 품질 평가를 위한 연구로는 서론에서 언급한 추천 수 또는 조회 수 등 서비스 제공자 측에서 수집하는 비텍스트 정보를 이용하여 지식문서의 품질을 평가하는 방법을 제안한 연구[1]를 들 수 있다. 이 방법의 경우 추천 수, 조회 수 및 답변 길이 등 지식문서의 답변과 관련된 총 13가지의 비텍스트 정보를 활용하여 기계학습 기반의 확률적 품질 평가 분류기를 만들어 품질지수를 측정하고, 이를 검색 모델에 반영하여 지식문서 검색 성능을 향상시켰다. 제안된 자질 중 가장 중요한 자질은 답변 길이였다. 그리고 길이를 제외한 나머지 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 하여도 이전에 작성된 답변에 비해 비텍스트 정보의 양이 부족할 수밖에 없고, 따라서 공평한 평가가 어려운 단점이 있다.

이전 연구에서 답변 품질 측정에 가장 유용하다고 나타난 자질이 길이 자질이라는 점에 착안하여 답변에 포함된 단어 개수나 문장 개수 및 문장 평균 길이 등 길이와 관련된 다양한 표면(surface) 자질을 활용하여 지식문서 품질 평가의 성능을 높인 연구[8]도 존재하였다. 하지만 제안된 표면 자질만을 사용할 경우 길이에 관련된 자질만으로 답변의 품질을 예측하기 때문에 답변 내용 자체의 적합성이나 신뢰성 등을 파악하기 어렵다는 단점이 있다.

본 연구에서는 기존 연구들에서 이미 문서 품질 평가에 유용하다고 증명된 길이 자질을 기본적인 신뢰도 자질로 놓고, 길이 자질과 더불어 사용했을 때 성능 보완 작용을 할 수 있는 여러 신뢰도 자질에 기반을 두는 지식문서 품질 평가 모델을 제안한다.

3. 지식문서 집합

본 연구에서 사용하는 실험 데이터 집합은 지식문서의 평가 기준을 제시한 기존 연구[7]에서 구축된 것으로서, 비텍스트 정보를 이용한 문서 품질 평가 연구[1]에서 사용된 것과도 동일한 집합이다. 본 데이터 집합은 2004년 6월 29일 하루 동안 네이버 검색엔진에 입력된 125개의 질의로 네이버 지식 질의응답 서비스에서 검색된 9,104개의 질문들 중에서 문헌정보학 전공자들에 의해 질의에 적합하다고 판정된 1,700개의 질문과 전체 네이버 지식문서 컬렉션에서 무작위로 선정된 894개의 질문, 그리고 각 질문에 대한 답변들 중 질문자가 "Best Answer"로 선택한 답변으로 이루어져 있으며, 각 답변에 대하여 신뢰도가 3점 척도(높음, 보통, 낮음)를 기준으로 문헌정보학 전공자들에 의해 평가되어 있다[7].

각 답변의 신뢰도 평가 과정에서는 질문의 유형에 따라 요구되는 답변의 수준이 다르기 때문에 먼저 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 다르게 적용되었다. 구체적인 답변의 신뢰도 평가 기준은 표 1과 같으며, 이 기준의 평가 항목 중 하나 이상을 만족시키면 기준에 부합되는 것으로 평가되었다[7]. 표 2 및 표 3은 데이터에 실제 존재하는 답변 중 신뢰도가 높음 또는 낮음으로 책정된 예를 하나씩 첨부한 것이다.

본 연구에서는 [1]에서 한 방법과 동일하게 무작위로 선택된 질문에 대한 답변들을 학습 집합으로 사용하고, 질문이 질의에 적합하다고 판정된 답변의 집합을 평가

표 1 답변의 신뢰도 평가 기준[7]

신뢰도 정보	지식형	생활형
높음	• 공신력 있는 정확한 출처 • 객관적으로 확실한 근거 (이론적, 학문적인 예시 등) • 논리적 설명 • 자료 첨가 (표, 그림, 사진) • 질문에 대한 핵심 답변	• 논리적인 개인의견 • 속담, 격언, 생활지식 등 학문적 근거는 없지만 상식적 • 질문에 대한 핵심 답변
보통	• 답변을 하였으나 근거부족	• 정확한 출처가 나오지는 않고 답변의 의견에 의존하나 어느 정도 논리적임
낮음	• 비방, 욕설, 음란한 글 • 명예훼손성 글 • 추측성 답변 • 근거가 없는 개인의견 • 질문과는 전혀 관련 없는 답변	

표 2 신뢰도가 높다고 판정된 답변의 예

질문 제목	바스티안같은 게임좀가르쳐주세욤
질문 본문	바스티안이 잼있었느데 유료라 하지 못하고있ㄴ욤...바스티안같은 그런 게임좀 소개 시켜주세욤..^^
답변 제목	re: 바스티안같은 게임좀가르쳐주세욤
답변 본문	바스티안을 하지 못하시구 계시는 이유가 '유료'때문이라면 바스티안을 다시 시작해 보시는게 어떨까요 ^^? 바스티안은 2004년 4월 8일 무료화 되었답니다 홈피 가보시면 아시겠지만 만약 섭이 바티아 가 아니시라면 바티아로 섭 이전 신청을 하셔야 할겁니다 ^^ *(이하 생략)*
신뢰도가 높은 이유	객관적으로 확실한 근거(날짜 제시), 질문에 대한 핵심 답변

표 3 신뢰도가 낮다고 판정된 답변의 예

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

표 4 데이터의 신뢰도 분포

집합	높음	보통	낮음
학습용	602 (67%)	213 (24%)	78 (9%)
평가용	1,099 (65%)	393 (23%)	208 (12%)
전 체	1,701 (66%)	606 (23%)	286 (11%)

집합으로 사용한다. 표 4는 학습용 집합과 평가용 집합의 신뢰도 분포를 나타낸다.

4. 신뢰도 자질 기반 품질 평가 모델

이 장에서는 본 연구에서 제안하는 지식문서의 품질 평가를 위한 신뢰도 자질 및 문서 품질 평가 모델에 대해 설명한다.

4.1 지식문서의 품질 평가를 위한 신뢰도 자질

지식 질의응답 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 어려울 정도로 불성실하게 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 자연어처리 기술을 이용하여 문서를 분석하는 과정이 필요하다고 볼 수도 있다. 하지만 본 연구에서는 지식 질의응답 서비스가 웹에서 제공되고, 따라서 많은 양의 문서가 품질 평가 대상이 될 수 있다는 점에 착안하여 컴퓨터가 계산하기 쉬운 변수들을 자질로 사용한다. 자질들 중 대부분은 어휘사전(lexicon) 기반으로 측정하게 된다. 즉, 높은 품질의 답변에서 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 자질은 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.
- **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.
- **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 신

뢰도를 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.
- **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 주소의 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.

연결어나 추정어 등을 위한 사전은 두 명의 학생이 신뢰도가 높은 답변에서 주로 나오는 단어 목록과 신뢰도가 낮은 답변에서 가장 빈번하게 출현하는 단어 목록에서 직접 뽑아서 구축하였다. 사전에 기록될 단어를 뽑을 때에는 두 명의 동의하에서만 뽑았다. 그리고 이모티콘의 경우, 광고성 문자메시지에 주로 출연하는 어구 위주로 사전을 구축하였다. 각 사전기반 자질의 어휘사전

표 5 사전기반 자질의 예제

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

표 6 각 어휘사전의 크기

자질	사전 크기
연결어	69
가치판단어	44
추정어	45
개인의견 단어	30
광고성 단어	25
이모티콘	1,559

표 7 각 자질과 신뢰도 점수의 Pearson 상관계수

자질	상관계수
연결어	+0.0127
가치판단어	-0.1323
추정어	-0.2092
개인의견 단어	-0.1974
광고성 단어	-0.0278
이모티콘	-0.2777
길이	+0.2407

에 등재된 단어의 예는 표 5에 나타나 있다.

표 6은 각 사전기반 자질을 위해 구축된 각 사전에 등재된 단어 또는 구의 개수, 즉 사전의 크기를 나타낸다.

표 7은 제안된 각 신뢰도 자질과 학습용 집합 데이터의 신뢰도 점수(높음=2, 보통=1, 낮음=0)의 Pearson 상관계수를 나타낸다.[1] Pearson 상관계수는 다음과 같이 계산된다.

$$corr = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2(y-\bar{y})^2}}$$

위 식에서 x는 문서에서 추출된 자질의 값이고, y는 신뢰도 점수를 의미한다. 상관계수의 값이 양수이면 자질의 값이 커질수록 신뢰도 점수도 높아진다는 것을 나타내고, 음수이면 자질의 값이 작아질수록 신뢰도 점수가 높아진다는 것을 나타낸다. 상관계수가 0에 가까울수록 실제 신뢰도 점수와의 관련성이 낮음을 나타낸다.

상관계수 측정 결과를 보게 되면 이모티콘 자질의 경우 기존에 이미 유용하다고 알려진 길이 자질보다 더 높은 상관계수가 나온 것을 알 수 있다. 연결어와 광고성 단어의 경우 직관과 달리 실제 신뢰도 점수와의 관련성이 매우 낮음을 알 수 있다. 가치판단어의 경우 직관과 달리 상관관계가 음수로 나온 것을 알 수 있다.

4.2 문서 품질 평가 모델

본 연구에서는 비텍스트 정보를 다룬 기존 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. 본 연구에서는 Zhang Le의 최대 엔트로피 모델화 툴킷[2]을 사용하여 품질 평가 모델을 구현하였다.

주어진 답변을 문서 x라고 하고 이 x라는 문서에 매길 수 있는 품질 등급을 $y=\{$높음,보통,낮음$\}$라고 하자. 본 품질 평가 모델의 목적은 조건부 확률 $p(y=$높음$|x)$

1) 웹사이트 주소의 출현 여부 자질은 이진 자질이므로 제외하였다.
2) http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

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 i번째 자질의 값을 출력하는 함수이고, λ_i는 학습 과정에서 Limited-Memory Variable Metric 기법[10]을 이용하여 미리 계산된 i번째 자질 함수의 가중치이며, Z는 모든 가능한 y에 대하여 $p(y|x)$을 더했을 때의 값이 1이 되도록 하기 위한 정규화 인수이다.

최대 엔트로피 기반 모델은 다양한 정보를 자질로 결합할 수 있는 장점을 가지면서 조건부 확률을 추정할 수 있다. 본 품질 평가 모델이 출력하는 확률은 추후 언어 모형(language model) 기반 검색모델[11]에 적용하여 검색 성능을 향상시킬 수 있다.

5. 실험 및 평가

이 장에서는 본 연구에서 제안하는 신뢰도 자질 기반 문서 품질 평가 모델의 실험과 성능에 대해 설명한다.

5.1 평가 척도

본 연구에서는 품질 평가 모델이 출력하는 조건부 확률 즉, 주어진 답변의 신뢰도가 높을 확률로 순위가 매겨진 답변 목록을 대상으로 기존 정보검색 시스템 평가에 전통적으로 사용되는 평균 정확률(average precision) 및 일반화된 정확률과 재현율(generalized precision and recall)[12]로 모델의 성능을 측정한다. 다양한 평가 척도를 사용하는 이유는 평가의 신뢰성을 좀 더 높이기 위함이다.

먼저 평균 정확률 AP은 다음과 같이 측정한다.

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

위 수식에서 r은 순위, N은 검색된 문서의 개수(본 실험에서는 전체 답변의 개수), $P(r)$은 r번째 순위까지의 정확률, 그리고 $isrel(r)$은 r번째 순위의 문서의 적합 여부(본 실험에서는 신뢰도가 높거나 보통일 경우 1, 낮을 경우 0)를 출력하는 함수를 의미한다. 그리고 본 실험에서는 신뢰도가 높거나 보통인 답변을 적합한 문서라고 간주한다.

일반화된 정확률과 재현율은 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 정확률 gP와 재현율 gR의 정의는 다음과 같다.

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

여기에서 $r(d)$는 문서 d의 점수, R은 집합 $D=\{d_1,d_2,...,d_n\}$에서 검색된 문서 집합(본 연구에서는 전체 답변), 그리고 N은 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. 일반화된 정확률을 이용하면 일반화된 평균 정확률 AgP도 다음과 같이 계산할 수 있다.

$$AgP=\frac{\sum_{r=1}^{N}(gP(r)\times isrel(r))}{\text{적합한 문서의 개수}}$$

요약하자면 AP는 신뢰도가 높거나 보통인 답변을 같은 등급으로 묶어 성능을 평가하는 것이고, AgP는 세 등급(높음, 보통, 낮음)을 모두 고려하여 평가하는 방식이라 할 수 있다.

5.2 실험 결과 및 분석

먼저 제안하는 각 신뢰도 자질의 상대적인 문서 품질 평가 성능 기여도를 알아보기 위해 모든 신뢰도 자질 조합에서 해당 자질만 제외한 상태의 품질 평가 모델의 성능을 AP 및 AgP로 측정하였다. 실험의 결과는 표 8에 나타나있다. (표에서 "ALL"은 모든 신뢰도 자질을 의미한다.)

표 8의 결과로 각 자질을 분석해보면 길이 자질이 타 자질들에 비해 가장 기여도가 높다는 것을 알 수 있다. 나머지 자질들은 길이 자질만큼의 영향력은 갖지 않으나 전체적인 성능을 보완해주는 역할을 한다는 것을 표 8의 결과에서 암시한다고 볼 수 있다. 이전에 보았던 표

표 8 각 자질 제외 시 성능 변화

자질	AP (%)	AgP (%)
ALL	97.45	91.77
ALL - 연결어	97.48(△0.03)	91.74(▽0.03)
ALL - 가치판단어	97.48(△0.03)	91.58(▽0.19)
ALL - 추정어	97.33(▽0.12)	91.53(▽0.24)
ALL - 개인의견 단어	97.45(−0.00)	91.65(▽0.12)
ALL - 광고성 단어	97.45(−0.00)	91.76(▽0.01)
ALL - 이모티콘	**97.10(▽0.35)**	**91.44(▽0.33)**
ALL - 웹사이트 주소	97.46(△0.01)	91.77(−0.00)
ALL - 길이	**92.10(▽5.35)**	**82.01(▽9.76)**

7에서 길이 자질과 동일하거나 더 높은 학습데이터와의 상관계수를 보인 이모티콘 자질과 추정어 자질 등 학습 데이터를 이용하여 구축된 사전에 기반하여 추출된 몇몇 자질들의 경우, 평가데이터에 대한 적용 범위(coverage)가 제한적이어서 실제 문서 품질 평가에 대한 성능 기여도가 상관계수만큼 크지 않게 나온 것이라 판단된다. 추후 연구에서는 사전기반 자질 추출에 사용되는 사전들을 확장하여 적용 범위를 넓히는 방법을 모색하고자 한다.

앞서 기존 연구를 살펴본 바와 같이, 가장 최근의 지식문서 품질 평가 방법은 비텍스트 정보를 이용한 방법[1]이 대표적이다. 본 실험 환경의 지식문서 데이터와 품질 평가 모델은 모두 [1]에서 사용된 것과 동일하기 때문에 비텍스트 자질과 제안하는 신뢰도 자질 간의 성능 비교가 가능하다.

제안하는 신뢰도 자질과 기존 비텍스트 자질의 성능 비교 실험 결과는 그림 1 및 표 9와 같다. (그림 1에서 "Text-Confidence"는 제안하는 신뢰도 자질 조합을 의미하고, "Non-textual"은 비텍스트 자질 조합을 의미한다.) 본 연구에서 비텍스트 자질로 사용한 정보는 답변 길이, 조회 수, 답변 수, 사용자 추천 수, 사용자 비추천 수, 출력회수, 스크랩회수, 답변자 채택률, 질문자 부여 평점, 답변자 내공, 답변자 전문가 지수, 관리자 추천 여부 및 스폰서 답변 여부 등 13가지로 기존 비텍스트 자질에 관한 연구[1]에서 사용한 것과 동일하다.

위 실험 결과에서 지식문서 품질 평가에 있어서 본 논문에서 제안하는 신뢰도 자질이 비텍스트 자질에 비해

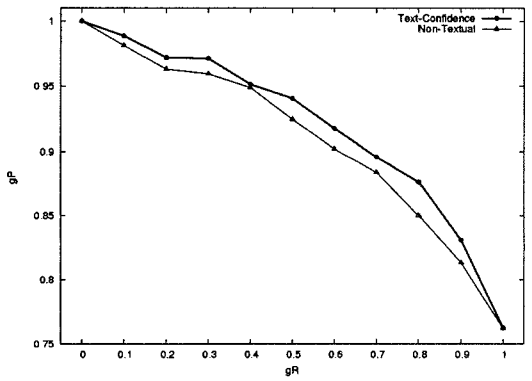

그림 1 gR-gP 그래프로 측정한 기존 비텍스트 자질[1]과의 성능 비교

표 9 AP 및 AgP로 측정한 두 접근법의 성능

자질 조합	AP (%)	AgP (%)
비텍스트[1]	97.01	90.76
신뢰도	**97.45**	**91.77**

표 10 길이 자질 제외 시 두 접근법의 성능 비교

자질 조합	AP (%)	AgP (%)
길이	96.80	90.38
비텍스트[1] - 길이	92.42	83.14
신뢰도 - 길이	92.10	82.01

성능을 약 1% 정도 개선하였음을 알 수 있다. 주목할 점은 기존 방법[1]에서 답변 길이 외에도 조회 수, 추천 수 등 서비스 사용자의 참여에 의해 구축된 정보를 추가적으로 사용한 반면에, 본 실험에서 사용한 신뢰도 자질은 주어진 하나의 품질 평가 대상 문서의 텍스트에서만 추출한 것이기 때문에 활용하는 정보의 양이 다르다는 점이다. 제안하는 방법이 비교적 적은 양의 정보만을 자질로 이용하였음에도 불구하고 기존 방법과 대등한 성능을 보였다는 것은 제안하는 신뢰도 자질의 효율성과 유용성을 입증하는 것이라 볼 수 있다. 추후 연구에서는 보다 더 효과적인 신뢰도 자질 추출 방법을 모색하여, 기존 방법 대비 성능의 개선 폭을 향상시키는 데에도 주력할 예정이다.

표 10은 두 접근법 모두 길이 자질에 많은 부분 의존한다는 점에 착안하여 길이를 제외한 나머지 자질들로 품질 평가를 수행했을 때의 성능 비교이다. 길이 자질을 제외했을 때에는 신뢰도 자질이 비텍스트 자질에 비해 낮은 성능을 보이긴 하지만, 성능 변화폭이 크지는 않다. 추후 연구에서는 보다 더 정교한 신뢰도 자질 추출 방법을 모색하여 성능을 개선시킬 계획이다.

6. 결 론

본 논문에서는 사용자의 참여에 의해 작성된 지식문서의 품질을 평가하기 위한 자질로서 자료부족 문제에 견고한 신뢰도 자질을 제안하였다. 또한 네이버 지식 질의응답 서비스에서 수집한 실제 문서들을 대상으로 하여 확률 기반의 문서 품질 평가 모델에 제안한 자질을 적용하였고, 실험 결과를 통해 신뢰도 자질이 지식문서의 품질 평가에 있어 유용함을 입증하였다. 제안하는 방법이 기존 방법에 비해 상대적으로 적은 양의 정보를 사용하였음에도 불구하고 대등한 성능을 보인 것을 고려할 때, 상당히 의미 있는 성능 향상으로 판단된다.

본 연구에서 제안한 신뢰도 자질은 지식 검색 분야 문서에 종속적인 특징에 기반을 둔 것이 아니기 때문에, 지식 질의응답 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 기반을 둔 품질 평가 모델의 성능을 개선하기 위한 방안 및 모델의 출력결과를 실제 검색 모델에 반영하여 일반적인 사용자

제작 문서의 검색 성능을 효과적으로 향상시킬 방안에 대해 연구하고자 한다.

참 고 문 헌

[1] Jeon, J., Croft, W. B., Lee, J. H., and Park, S., "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] Page, E. B., "Computer grading of student prose, using modern concepts and software," Journal of Experimental Education, Vol.62, No.2, pp.127-142, 1994.

[3] Brin, S. and Page, L., "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] Kleinberg, J. M., "Authoritative sources in a hyperlinked environment," J. ACM, Vol.46, No.5, pp.604-632, 1999.

[5] Zhu, X. and Gauch, S., "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] Zhou, Y. and Croft, W. B., "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, pp.299-314, 2006.

[8] Lee, J.-T., Song, Y.-I., and Rim, H.-C., "Predicting the quality of answers using surface linguistic features," In Proceedings of the 6th International Conference on Advanced Language Processing and Web Information Technology, pp. 111-116, 2007.

[9] Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D., "A maximum entropy approach to natural language processing," Comput. Linguist. Vol.22, No.1, pp.39-71, 1996.

[10] Malouf, R., "A comparison of algorithms for maximum entropy parameter estimation," In Proceedings of the Conference on Computational Natural Language Learning, pp.49-55, 2002.

[11] Ponte, J. M. and Croft, W. B., "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] 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.

이 정 태

2006년 고려대학교 컴퓨터학과 학사. 2008년 고려대학교 컴퓨터학과 석사. 현재 고려대학교 컴퓨터·전파통신공학과 박사과정. 관심분야는 정보검색, 자연어처리

송 영 인

2001년 고려대학교 컴퓨터학과 학사. 2003년 고려대학교 컴퓨터학과 석사. 2008년 고려대학교 컴퓨터학과 박사. 관심분야는 정보검색, 자연어처리

박 소 영

1997년 상명대학교 전자계산학과 학사. 1999년 고려대학교 컴퓨터학과 석사. 2005년 고려대학교 컴퓨터학과 박사. 2007년~현재 상명대학교 디지털미디어학부 교수. 관심분야는 자연어처리, 기계번역, 기계학습

임 해 창

1981년 University of Missouri-Columbia 전산학과 학사. 1983년 University of Missouri-Columbia 전산학과 석사. 1990년 University of Texas at Austin 전산학과 박사. 1991년~현재 고려대학교 컴퓨터·통신공학부 교수. 관심분야는 자연어처리, 정보검색, 한국어정보처리