질의응답시스템을 위한 문서의 품질 평가[1)]

이형규*O 김민정* 신중휘* 이정태* 윤여찬** 임해창*
고려대학교 컴퓨터·전파통신공학과*
한국전자통신연구원**
{hglee, mjkim, jhshin, jtlee, rim}@nlp.korea.ac.kr*, ycyoon@etri.re.kr**

Document Quality Evaluation for Question Answering System

Hyoung-Gyu Lee*O Min-Jeong Kim* Joong-Hwi Shin*
Jung-Tae Lee* Yeo-Chan Yoon** Hae-Chang Rim*
Dept. of Computer and Radio Communications Engineering, Korea University*
Electronics and Telecommunications Research Institute**

요 약

본 논문에서는 질의응답시스템에서 응답 추출 대상 문서로 사용할 적절한 문서를 찾는 방법으로 기계 학습 기반의 문서 품질 평가 기법을 사용한다. 본 논문에서는 기존 연구와 달리 객관적인 정보를 많이 포함하고 있는 문서를 선별하는 목적으로 문서 품질 평가를 위한 유용한 자질들을 제안한다. 본 논문에서 정의한 정보성 자질은 정보의 양을 측정하는 자질과 정보의 객관성을 측정하는 자질로 구성된다. 실험 결과, 기존 문서 품질 평가 연구에서 주로 사용된 자질들만 사용한 경우와 새로운 자질들을 추가한 경우를 비교하였을 때, 1.5배 정도 높은 평균 정확률을 보였다. 제안하는 자질들 중에는 정보성 자질이 매우 유용한 자질이었고, 가독성 자질은 비교적 낮은 성능을 보였다. 문서의 여과 실험 결과, 96.4%의 재현율을 유지하면서 전체 문서 집합 중, 60%에 해당하는 저품질 문서를 여과할 수 있었다.

1. 서론

질의응답시스템은 사용자가 입력한 질문에 대해 문서 집합으로부터 적절한 응답을 자동으로 찾아주는 시스템이다. 최근 질의응답시스템은 웹 문서를 응답 추출 대상 문서로 사용하고 있다[1,2]. 웹 상에는 새로운 문서가 빠르게 생산되므로 사용자가 필요로 하는 정보를 포함하는 문서가 많기 때문이다.

그러나 웹 상의 모든 문서를 질의응답시스템에서 응답 추출 대상 문서로 사용하기에는 그 문서의 수가 많아 응답 속도의 저하를 가져온다는 문제가 있다. 따라서 질의응답시스템을 위한 문서 집합으로 사용하기에 적절치 않은 문서를 전처리 단계에서 여과하는 작업이 필수적이다. 이와 같이 응답 추출 대상 문서로서 가치가 높은 문서만을 대상으로 문서 색인 및 응답 추출을 수행한다면 질의응답시스템의 속도를 향상시킬 수 있다.

<그림 1>과 <그림 2>는 질의응답시스템 측면에서 고품질 문서와 저품질 문서의 예를 보여준다. <그림 1>은 게임의 규칙을 설명하고 있는 문서이므로 질의응답시스템 측면에 필요한 정보가 많은 문서라고 할 수 있다. 이러한 문서는 질의응답시스템 측면에서 우선적으로 응답 추출 대상이 되어야 할 문서이다. <그림 2>는 저자의 개인적인 느낌과 의문점을 서술한 문서로 질의응답시스템이 사용할 정보가 거의 없는 문서라고 할 수 있다. 이러한 문서는 사전에 여과되어야 할 문서이다. 따라서 <그림 2>의 문서를 미리 여과할 수 있으면 시스템 속도를 향상시킬 수 있다.

이를 위해, 주어진 문서가 질의응답시스템에서 응답 추출 대상으로서 적절한지를 평가하기 위한 방법이 요구된다. 본 논문에서는 그 방법으로 문서 품질 평가 방법을 사용한다. 기존에 알려진 문서 품질 평가 방법들[3-7]은 텍스트 자질 및 비텍스트 자질을 통해 문서의 품질을 다양한 관점에서 평가할 수 있고, 문서의 품질 수준을 점수화 해준다. 따라서 적절한 임계값을 찾아내면 응답 추출 대상으로 적절하지 않은 문서를 여과하기 위한 목적으로 사용하기에 적합한 방법이다.

기존의 문서 품질 평가에 관한 연구들은 대부분 정보검색의 성능 향상이라는 관점에서 이루어졌다. 정보검색시스템은 질의에 대해 적합한 문서 자체를 찾아주는 시스템이기 때문에 문서 품질을 평가하는 자질로서 저자의 신뢰도가 반영되는 비텍스트 자질 및 글의 신뢰도가 반영되는 텍스트 자질 등 다양한 자질들을 고려하였다. 그러나 정보검색시스템은 질의에 적합한 문서를 찾아야 하는 반면에 질의응답시스템은 문서 내에서 응답을 추출해야 한다. 따라서 문서 내에 응답으로 사용가능한 정보의 유무나 정보의 객관성 측면이 특히 중요하다. 따라서 질의응답시스템을 위한 문서의 품질 평가는 정보성 관점에서 중점적으로 평가할 수 있는 모델을 사용하여야 한다. 본 논문에서는 문서의 정보성 자질들을 정의하고 제안하는 자질들이 질의응답시스템 측면에서 문서 품질 평가에 유용한 자질임을 보인다. 본 논문은 질의응답시스템의 효율성 향상을 위해 문서 품질 평가를 시도한 연구로서

1) 본 연구는 지식경제부 및 정보통신연구진흥원의 IT신성장동력핵심기술개발사업의 일환으로 수행하였음.[A1100-0801-1408, 웹 QA 기술 개발]

.... 방장의 위대한 권한 이건 모르시는 분들이 조금 많을지도 모릅니다. 그리고 첫라운드에만 가능한 겁니다. 서든의 게임중 참여 가능한 특성상 스포나 다른 게임처럼 맵에 들어간 후 일정 시간 기다리면서 유저들이 다 정상적으로 접속한 후 게임이 시작 되는 것이 아닙니다. 방장이 게임시작 버튼을 누르고 맵 로딩후 방장이 게임을 시작할수 있는 상태가 되면 다른유저를 기다리는일 없이 바로 플레이 되는 게 서든입니다. 컴 사양과는 상관 없이 방장이 시작한 이후에만 다른 유저들도 플레이 가 가능합니다.. 이해 안가시죠? 쉽게 말해서 클전을 제외한 일반 공방전에서는 첫라운드는 무조건 그 누구도 방장보다 먼저 움직일수 없다는 이야기죠. 다르게 말하면 방장은 첫라운드만은 아군과 적군들보다 2초에서 5초정도 먼저 움직일수(게임을 플레이 할수) 있다는 말입니다 ....

<그림 1> 질의응답시스템 측면에서 고품질 문서의 예

.... 격노너프랑 이속 평준화가 아주 문제군요-ㅁ- 특히 이속평준화는 좀..... 그래도 치타상도 이제 도트에 멍해지지 않고 뭣보다도 펫스킬 다양해진게 기쁩니다!ㅋㅋ 천둥매의 천둥숨결이라는게 있다고 하던데 옛날부터 그 스킬 한번 쓰게 하고싶다 이러고 있었는데+ + 한번 키워보고 싶은.. 이제 서포트형 펫의 시대가 열리는것인가[두둥] 아 그리고 루포스 너프 안됐나요? 패치내용에서는 못봤는데 ....

<그림 2> 질의응답시스템 측면에서 저품질 문서의 예

의미가 있다. 제안하는 자질을 사용한 품질 평가 모델은 문서를 점수화하여 적절한 문서만을 응답 추출 대상으로 삼을 수 있도록 해준다. 이는 웹 문서를 대상으로 하는 질의응답시스템의 속도를 개선할 수 있다는 점에서도 의미가 있다.

2. 기존 연구

질의응답시스템을 위한 문서의 품질 평가를 시도한 기존 연구는 없으나, 정보검색시스템의 측면에서 웹 문서의 여과와 관련된 연구와 문서 품질 평가에 관련된 연구는 다수 존재한다.

Liu의 연구[8]에서는 정보검색 대상으로 가치가 낮은 문서를 여과하기 위해 K-means 클러스터링에 기반한 핵심 리소스 선택 방법(key resource selection)을 제안하였으며 정보검색에 적용하여 유용성을 입증하였다.

Zhu의 연구[3]에서는 문서의 품질을 평가하기 위해 통용성, 이용가능성, 정보 대비 노이즈 비율, 저자의 신뢰성, 인기도 등 다양한 관점에서 문서를 평가하였다. 이 연구에서는 정보 대비 노이즈 비율을 통해 정보성 관점에서 문서를 평가하고자 시도하였으나, 단순히 문서 내의 텍스트의 비율을 자질로 사용하였다.

다른 기존 연구들[4-7]에서는 cQA[2] 문서를 평가하기 위해 질문의 특징, 답변의 특징, 질문과 답변의 관계 등을 활용하였다. 또한 비텍스트 자질의 유용성이나 어휘 자질을 중심으로 한 텍스트 자질의 유용성을 연구하였다.

이러한 문서 품질 평가와 관련된 기존 연구들은 공통적으로 정보검색시스템의 성능 향상에 초점을 맞추어 수행되었다. 질의에 적합한 문서의 검색이 목표인 정보검색시스템 측면에서는 저자의 신뢰도를 반영하는 각종 비텍스트 자질들[4,5,7]과 글의 신뢰도를 반영하는 각종 텍스트 자질들[6] 그리고 가독성 자질들[5,6]이 유용하였다. 그러나 질의응답시스템과 같이 문서 내에서 정보를 추출하여 출력해주는 시스템의 측면에서는 문서 품질의 기준이 다르다. 따라서 정보검색시스템 측면에서 연구되었던 품질 평가 방법과 같은 기준을 질의응답시스템에 적용하는 것은 적절하지 않다. 이에 본 논문에서는 정보의 객관성에 초점을 맞춘 다음과 같은 자질들을 사용할 것을 제안한다.

3. 정보성 자질

객관적인 정보를 많이 갖고 있는 문서가 질의응답시스템이 찾고자 하는 정답을 포함할 가능성이 높다는 가정하에 정보성 자질을 정의한다.

본 논문에서 사용한 정보성 자질은 정보의 양을 평가하는 자질, 객관성을 평가하는 자질, 특정 도메인에 종속적인 자질로 구성된다. 본 논문에서는 기존 문서 품질 평가에서 사용한 일부 자질을 정보성을 평가하기 위한 자질로 새롭게 해석하여 정보성 자질에 포함시켰다.

3.1 정보의 양을 평가하는 자질들

문서 내에 포함된 정보의 양은 정보성을 파악하기 위한 필수 요소 중 하나이다. 정보의 양을 평가하기 위해 다음과 같은 자질을 사용한다.

· 길이 기존 연구에서 문서의 길이가 문서의 품질을 평가하는 데에 좋은 자질이라고 알려져 있다. 본 논문에서도 역시 문서의 길이가 길수록 정보를 많이 포함할 수 있다고 가정한다. 문서 본문의 길이를 byte 단위로 측정한 값을 하나의 자질로 사용하였다[4-7].

· 어휘 개수 문서 내에서 사용한 어휘의 개수가 많을수록 정보가 많을 것이라고 가정한다. 문서 내에 출현한 어휘를 어절 단위로 분석하여 중복 제거한 어휘의 개수를 하나의 자질로 사용하였다[4-5].

· 이미지 및 동영상 첨부 여부 문서의 작성자는 독자의 이해를 돕기 위한 방법으로 이미지를 삽입한다. 독자는 글만으로는 얻을 수 없는 많은 정보를 이미지나 동영상을 통해 습득할 수 있다. 따라서 하나 이상의 이미지나 동영상이 포함된 문서는 이미지가 없는 문서보다 정보성이 높다고 가정한다.

3.2 정보의 객관성을 평가하는 자질들

양적으로 풍부한 문서라고 해서 그 문서가 반드시 정보성이 높다고 할 수는 없다. 일기와 같이 개인적인 생각이나 느낌을 주로 서술한 글은 다른 많은 사람들에게 보편적으로 유용한 정보가 적으므로 응답 추출 측면에서

2) 커뮤니티 기반 질의응답 서비스(Community based Q&A)를 말한다. Naver의 지식인이나 Yahoo! Answers가 대표적인 서비스이다.

좋지 않은 문서이다. 따라서 문서 내에 포함된 정보의 객관성이 같이 평가되어야 한다.

본 논문에서는 정보의 객관성을 평가하기 위한 자질로서 다음과 같은 다섯 범주의 어휘 자질들을 제안한다. <표 1>은 각 어휘 자질의 예를 보여준다.

· 설명문형 어휘 설명문형 어휘란 객관적으로 설명하는 글에서 많이 나타나는 어휘를 뜻한다. 설명문형 어휘를 많이 사용한 문서가 객관성이 높다고 가정한다.

· 개인성 어휘 일기나 개인 신상 관련 글에 자주 나타나는 어휘를 뜻한다. 개인적인 글은 질의응답시스템에 도움이 되는 정보를 포함할 가능성이 낮다고 가정한다.

· 추정 및 의견 어휘 추측이나 개인 의견을 표현할 때 많이 사용하는 어휘이다. 개인적인 의견이나 주장을 나타내는 어휘를 많이 사용한 문서는 주관적인 문서이며 이러한 문서는 정보성이 낮다고 가정한다[6].

· 가치 판단 어휘 가치 판단 어휘는 기존 연구에서 글의 성실성을 반영하는 자질로 제안되었다[6].

· 광고성 어휘 광고성 문서 내의 정보는 질의응답시스템에서 필요로 하는 객관적인 정보와는 거리가 있다[6].

본 논문에서 사용한 어휘 자질들은 어절 단위로 분석하였으며, 학습 집합의 고품질 문서와 저품질 문서 내의 어휘 분포를 조사하여 통계적으로 의미 있는 어휘를 선별하였다. 또한 고유 명사와 같이 문서의 품질과는 관계없이 특정 문서에 국한되어 나타나는 어휘를 수동으로 제거한 후, 남은 어휘로 사전으로 구축하였다. [6]에서는 어휘 자질의 선별 시, 전적으로 수동에 의존하는 방법을 사용하였으나, 본 논문에서는 통계적 접근법을 이용하여 반자동으로 구축한 특징이 있다. 통계적 접근법은 유용한 어휘 자질을 빠르게 추출할 수 있다는 장점이 있다. 통계 분석 기법으로는 카이제곱 검정 방법[9]을 사용하였으며, 수식은 다음과 같다.

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

O_i는 각 문서 집합에서 실제 출현 회수이고, E_i는 각 문서 집합에서의 출현 기대값이다. 카이제곱 점수가 높을수록 각 문서 집합의 특성을 나타내는 의미있는 어휘라고 할 수 있다. 사전 내 각 어휘를 자질로 사용할 때는 문서 내 출현 회수를 문서의 길이로 정규화한 값을 사용하였다.

3.3 도메인 종속적인 자질들

많은 질의응답시스템은 응답의 정확도를 높이기 위해 도메인을 한정하는 정책을 취한다. 따라서 도메인을 한정한 질의응답시스템을 위한 문서 품질 평가에서는 해당 도메인과 관련된 자질을 사용할 수 있다. 본 논문에서는 게임 도메인 말뭉치를 실험 데이터로 사용하였고 게임 도메인에 종속적인 <표 2>와 같은 자질들을 추가하였다.[3] 게임 유형, 게임 장르 또는 각 게임 별로 선호하는

<표 1> 정보의 객관성을 평가하는 어휘 자질

자질	예제
설명문형 어휘	있다, 한다, 등이, 있습니다, 통해, 비해
개인성 어휘	내가, 제가, 나는, 나의
추정 및 의견 어휘	것 같다, 것 같습니다, 제 생각에는, 생각한다
가치 판단 어휘	매우, 항상, 반드시, 최대한
광고성 어휘	무료, 대출, 포커

<표 2> 도메인 종속적인 자질

자질	예제
게임 유형	온라인, 비디오, 모바일, PC
게임 장르	MMORPG, 전략시뮬레이션, FPS/액션
게임명	루니아전기, 신야구, 스타크래프트

사용자 집단이 다르고 그 집단마다 작성하는 글의 정보 수준에 차이가 있을 것이라고 가정하였다.

4. 정보성 외의 자질들

본 논문에서는 3장에서 정보성 자질로 정의한 자질들 외의 다음과 같은 자질들을 제안한다.

4.1 가독성

일반적으로 구성이 잘 되어 있고 사용자가 읽고 쉽게 이해할 수 있는 문서가 품질 좋은 문서라고 알려져 있다[5,6]. 질의응답시스템 측면에서도 역시 가독성이 좋은 문서가 품질이 좋은 문서라고 할 수 있다. 질의응답시스템은 문서 내에서 단어를 추출하는 경우뿐만 아니라 문장들을 그대로 답변으로 추출하는 경우도 많다. 이러한 경우 문장의 가독성이 곧 답변의 가독성으로 이어지기 때문에 질의응답시스템 측면에서도 가독성 자질이 유용할 수 있을 것이라고 가정한다. 따라서 본 논문에서는 정보성 자질에 추가로 가독성 자질의 사용을 제안한다. 가독성을 평가하는 세부 자질로는 다음과 같은 자질들을 제안한다. 이 중, 어휘 자질은 정보성 자질과 마찬가지 방법으로 추출하여 적용하였다.

· 어휘 밀도 (Lexical density) 어휘 밀도가 높은 문서는 다양한 어휘를 사용하여 독자가 읽고 이해하기 쉬운 문서이다. 어휘 밀도는 다음과 같은 수식으로 계산된다[5].

$$\text{어휘 밀도} = \frac{\text{문서 내에 사용된 어휘 개수}}{\text{문서 내의 어절 개수}}$$

3) 본 논문에서 제안하는 자질들을 게임 외의 다른 도메인에 적용할 경우, 도메인 종속적인 자질들은 제거하거나 적절히 변경하여 적용할 수 있다. 예를 들어, 음식 도메인이라면, 게임 장르를 음식 종류로, 게임명을 음식명으로 변경할 수 있다.

· 연결 어휘 *따라서*, *또한*, *그래서* 등과 같은 연결어들은 문장과 문장 간의 연결을 매끄럽게 하기 위해 사용되며, 문서의 가독성을 높여준다[6].

· 통신 언어 및 이모티콘 ㅋㅋ, ^^, ㅠㅠ 등의 통신 언어나 이모티콘이 많은 문서는 가독성이 낮다[6].

4.2 저자 및 출처

신뢰도 높은 사이트 내의 문서 또는 신뢰도 높은 저자가 작성한 문서는 품질이 높다고 알려져 있다[3-5]. 저자 및 출처의 신뢰성이 높은 문서가 질의응답시스템의 응답으로서 가치가 있는 정보를 포함할 것이라는 가정 하에 저자 및 출처 자질의 사용을 제안한다. 세부 자질은 다음과 같다.

· 저자 고품질 문서를 많이 작성하는 저자와 저품질 문서를 많이 작성하는 저자를 구별하기 위한 자질이다. 문서 수집 시 같이 수집한 문서의 저자 정보를 하나의 자질로 사용하였다.

· 출처 사이트 고품질 문서를 많이 포함하는 사이트와 저품질 문서를 많이 포함하는 사이트를 구별하기 위한 자질이다. 문서 수집 시 문서의 출처 사이트 정보를 같이 수집하여 하나의 자질로 사용하였다.

· 출처 사이트의 PageRank PageRank는 웹 문서 간 링크 분석을 통해 웹 문서의 중요도를 수치화하여 보여준다. 본 논문에서는 PageRank로 계산되는 문서 출처 사이트의 중요도가 질의응답 측면의 품질 평가에도 도움이 될 것이라는 가정 하에 문서의 출처 사이트의 PageRank 값을 하나의 자질로 사용하였다[10,11].[4]

· 게시판 이름 문서의 출처 게시판 이름 역시 문서의 품질 평가에 도움을 줄 수 있다. 예를 들어, 공지사항 게시판이나 팁 게시판 문서는 많은 정보를 담고 있는 문서일 가능성이 높고, 스크린샷 게시판이나 자료실 게시판 문서는 질의응답시스템의 응답 추출에 유용한 정보를 담고 있는 경우가 상대적으로 적다. 이 자질을 사용하기 위해서는 문서 수집 시 그 문서가 작성되어 있던 게시판 이름을 같이 추출하여야 한다.[5] 본 논문에서 사용한 게시판 이름은 *공지사항*, *소설*, *팁 게시판*, *유저 게시판*, *질문 게시판*, *맵 자료실* 등이 있다.

5. 실험 및 평가

5.1 실험 환경

5.1.1 실험 데이터

실험 데이터로 게임 도메인 웹 사이트의 문서를 수집하여 사용하였다. 각 문서의 등급은 질의응답시스템 측면에서 유용한 정보를 포함하고 있는지 여부를 가장 중요한 기준으로 한 <표 3>과 같은 지침에 따라 평가하였다. 등급 평가는 두 사람이 동일한 문서를 각각 평가한 후, 불일치건에 대해서 정련 작업을 거쳐 대다수 사람들이 신뢰할 수 있는 등급을 부여하였다. 전체 문서 집합의 구성은 <표 4>와 같으며, 학습 집합 3184건과 실험 집합 1700건으로 나누어 실험에 사용하였다.

<표 3> 문서 평가 지침

등급	평가 지침 요약
A	· 유용한 정보를 충분히 담고 있으며 읽기에 지장이 없는 문서
B	· 유용한 정보는 담고 있으나 읽기 어려운 문서 · 유용한 정보는 담고 있으나 정보의 양이 지나치게 적은(1~2문장) 문서
C	· 유용한 정보를 담고 있지 않은 문서 · 지나친 은어의 사용으로 일부 사용자를 제외하면 해석이 불가능한 문서 · 내용을 파악할 수 없을 정도로 문법, 어법이 틀린 문서 · 거짓 정보를 제공하는 문서 · 일회성 정보(광고 글, 상업적인 글, 홍보하는 글 등)로 구성된 문서

<표 4> 실험 데이터의 구성

문서 등급	문서 개수	비율 (%)
A	392	8.03
B	274	5.61
C	4218	86.36
합계	4884	100.00

5.1.2 학습 모델

본 논문에서는 제안하는 자질들의 유용성을 실험하기 위한 학습 모델로 Support Vector Machine을 사용하였다. 모델은 주어진 문서가 A등급 문서일 확률을 출력한다. 이는 문서 품질 평가 모델에 의해 실험 집합 내의 문서들을 품질 순으로 순위화하기 위함이다.

5.1.3 평가 척도

본 논문에서는 학습 모델이 실험 집합 내 문서를 순위화한 결과를 평가한다. 평가 척도로는 순위화 모델 평가에 흔히 사용하는 평균 정확률(Average Precision)을 사용하였다.

5.1.4 기저 성능

본 논문에서 제안하는 응답 추출 대상 문서로서의 품질 평가 방법은 기존에 연구된 바가 없기 때문에 cQA 문서 또는 사용자 리뷰 문서의 품질 평가에 관한 연구[4-7]에서 제안한 자질만을 사용한 성능을 기저 성능으로 하여 비교 실험하였다. <표 5>의 세부 자질 중 진하게 표시된 자질은 본 논문에서 새롭게 제안한 자질이며 그 외 자질로 기저 성능을 측정하였다.

4) PageRank 자질은 Google의 툴바 질의(toolbar query)를 사용하여 각 URL의 PageRank 값을 0에서 10 사이의 정수로 얻어내어 사용하였다.

5) 게시판 이름 자질은 각종 게시판에서 수집된 문서에 한하여 사용할 수 있다.

<표 5> 자질별 성능 비교 (평균 정확률, 단위: %)

자질 그룹		세부 자질	단일 자질 성능	그룹별 성능		
Random Ranking			7.8			
기저 성능 ({길이,어휘 개수,추정 및 의견 어휘,가치 판단 어휘,광고성 어휘,어휘 밀도,연결어,통신어,PageRank})			58.20			
정보성 자질	정보의 양	길이	21.77	27	77.38	83.41
		어휘 개수	30.07			
		이미지/동영상 첨부 여부	8.24			
	정보의 객관성	***설명문형 어휘***	63.34	75.99		
		개인성 어휘	14.15			
		추정 및 의견 어휘	8.69			
		가치 판단 어휘	29.54			
		광고성 어휘	8.24			
	도메인 종속	***게임 유형***	10.2	16.9		
		게임 장르	8.24			
		게임명	11.93			
정보성 외 자질	가독성	어휘 밀도	25.87	41.81	75.34	
		연결어	22.85			
		통신어	11.52			
	저자 및 출처	***저자***	26.71	78.74		
		출처 사이트	34.34			
		PageRank	10.2			
		게시판 이름	67.39			
최적 (전체 - {길이, 이미지/동영상, 광고성 어휘, 게임 장르})			83.59			

5.2 실험 결과

5.2.1 제안하는 자질들의 성능 비교 실험 결과

본 논문에서 제안한 자질이 질의응답시스템의 응답 추출 대상 문서로서 적절한 문서를 얼마나 잘 찾아내는지를 실험한 결과는 <표 5>와 같다. 각 수치는 학습 모델이 1700건의 실험 집합 문서를 순위화한 결과에 대한 A 등급 문서 검색 관점의 평균 정확률이다.

실험 결과, 정보성 자질을 사용하지 않은 성능(75.34%)에 비해 정보성 자질만 사용한 성능(77.38)이 2% 정도 더 높으며, 정보성 자질을 추가하여 모든 자질을 같이 사용했을 때 8% 정도 더 높은 83.41%의 평균 정확률을 기록하였다. 따라서 본 논문에서 제안한 정보성 자질이 응답 추출 대상 문서의 적절성 평가에 유용함을 알 수 있다. 자질 그룹 중에는 정보의 객관성과 저자 및 출처의 신뢰성이 가장 좋은 성능을 보였다.

또한 정보성 자질 중 정보의 양을 평가하는 자질 그룹만 사용했을 때 성능은 매우 낮았다. 이는 문서 내의 정보는 양 뿐만이 아닌 정보의 객관성이 뒷받침되어야 정보성 있는 문서라고 할 수 있다는 가정에 부합하는 결과이다.

게임 유형과 같은 도메인 종속적인 자질들은 미미한 성능을 보였지만 정보성을 평가하는 자질들에 포함되어 사용하였을 때, 전체적인 성능을 조금 올려주는 효과를 보인다. 도메인 종속적인 자질들은 도메인에 따라 더 유용한 자질이 될 수도 있다.

단일 자질로서는 정보성 자질 그룹의 설명문형 어휘 자질과 저자 및 출처 자질 그룹의 게시판 이름 자질이 가장 높은 성능을 보였다. 이는 설명문형 어휘 자질이 객관적인 문서를 잘 찾아내어 질의응답 대상 문서의 적절성 평가에 큰 기여를 하고 있음을 보여주며, 정보의 객관성을 평가할 때 대표적인 자질로서 사용해야 함을 보여준다. 또한 게시판 이름 자질은 문서의 목적을 대체로 잘 반영하고 있기 때문에 매우 유용한 자질이라고 분석된다.

[6]에서 문서 품질 평가에 유용하다고 알려진 자질 중 광고성 어휘 자질은 성능 기여가 거의 없는 결과를 보였다. 이는 실험 데이터가 수집되기 전에 사이트 관리자가 광고성 글을 여과했기 때문으로 추정된다.

본 논문의 실험 결과와 기존 문서 품질 평가 연구와의 큰 차이는 [4-7]에서는 문서 길이 자질이 다른 자질들과 비교했을 때 매우 유용한 자질이었는데, 본 논문에서는 길이 자질이 성능 향상에 기여하고 있지 못하다는 점이다. 질의응답시스템에서 응답 추출 대상 문서로 적절한 문서는 길이보다는 정보의 객관성에 더 의존적이기 때문이다.

5.2.2 문서 여과 실험 결과

제안하는 자질들을 사용한 문서 품질 평가 방법이 질의응답시스템의 효율성 개선에 기여하는 정도를 알아보기 위해, 문서 집합의 여과 실험을 수행하였다. <표 6>은 위 실험 결과에서 알아낸 최적 자질 조합을 통해 평가된 테스트 집합 내의 문서들을 하위에서부터 일정량씩

여과하였을 때, A등급 문서의 재현율의 변화를 측정한 결과이다. 하위 60%의 문서를 여과하였음에도 96% 이

<표 6> 문서 여과량에 따른 A등급 문서의 재현율/정확률 변화

여과량	여과율	A등급 문서 재현율	A등급 문서 정확률
0	0%	100%	8.23%
340	20%	98.57%	10.15%
680	40%	98.57%	13.52%
1020	60%	96.43%	19.85%
1360	80%	92.14%	37.94%
1530	90%	79.29%	65.29%
1700	100%	0%	0%

상의 높은 A등급 문서의 재현율을 보였다. 이는 제안하는 방법이 응답 추출 대상으로 사용될 문서 집합의 크기를 줄여줌으로써 속도를 개선할 수 있음을 간접적으로 보여주는 결과라 할 수 있다.

6. 결론 및 향후 연구

본 논문에서는 질의응답시스템 측면에서 응답 추출 대상 문서의 적절성을 평가하기 위해 문서 품질 평가 방법을 사용하였다. 본 논문에서 정의한 정보성 자질은 기존 연구에서 사용한 자질들과 본 논문에서 새롭게 제안하는 자질들을 포함한다. 정보성 외에 가독성, 저자 및 출처 자질과 같은 기존에 연구되었던 자질들이 질의응답시스템 측면에서 문서 품질 평가에 왜 유용한지를 설명하였다.

제안하는 자질 중 정보성 자질을 중점적으로 고려하는 것이 질의응답시스템이 요구하는 고품질 문서를 찾는데 도움이 되었고, 정보성 자질 중에는 정보의 양을 평가하는 자질들보다는 정보의 객관성을 평가하는 자질들이 문서 품질 평가의 성능 향상에 더 많이 기여함을 보였다.

본 논문은 질의응답시스템의 효율성 향상을 위해 문서 품질 평가를 시도한 연구로서 의미가 있다. 또한 제안하는 자질을 사용한 품질 평가 모델이 질의응답시스템 측면에서 문서를 순위화하는 데에 유용함을 보였다. 이러한 순위화는 적절한 임계값을 통해 저품질 문서를 응답 추출 대상 문서 집합에서 사전에 제거할 수 있게 해줌으로써 질의응답시스템의 응답 시간을 단축할 수 있는 가능성을 제공한다.

향후 연구로서 질의응답시스템에 품질 평가 모델을 실제로 적용하여 질의응답 속도를 효과적으로 향상시킬 수 있는 방안의 연구가 필요하다. 그리고 질의응답시스템의 응답 정확도 측면에서 문서 품질 점수의 적용가능성을 연구할 필요가 있다. 또한 본 논문에서는 어휘 자질을 위해 사용될 사전을 반자동으로 구축하였는데, 향후에는 문서 품질 평가를 위한 어휘 사전을 자동으로 구축하는 방법의 연구가 필요하다.

참고 문헌

[1] C. Clarke, G. Cormack and T. Lynam, Web Reinforced Question Answering, In Proceedings of the Tenth TREC, 2001.

[2] C. Kwok, O. Etzioni and D. Weld, Scaling Question Answering to the Web, In Proceedings of the 10th WWW, pp. 150-161, 2001.

[3] X. Zhu and S. Gauch, Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web, In Proceedings of the 23rd ACM SIGIR, pp. 288-295, 2000.

[4] E. Agichtein, C. Castillo, D. Donato, A. Gionis and G. Mishne, Finding High-Quality Content in Social Media, In Proceedings of the 1st ACM WSDM, pp. 183-194, 2008.

[5] L. Hoang, J.-T. Lee, Y.-I. Song and H.-C. Rim, A Model for Evaluating the Quality of User-Created Documents, In Proceedings of the 4th AIRS, pp. 504-509, 2008.

[6] 이정태, 송영인, 임해창, 신뢰도 자질을 이용한 지식 검색 문서의 품질 평가, 제19회 한글 및 한국어 정보처리 학술대회 논문집, 62-67쪽, 2007년.

[7] 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 ACM SIGIR, pp. 228-235, 2006.

[8] Y. Liu, C. Wang, M. Zhang and S. Ma, Web Data Cleansing for Information Retrieval using Key Resource Page Selection, Special interest Tracks and Posters of the 14th WWW, pp. 1136-1137, 2005.

[9] E. W. Weisstein, Chi-Squared Test, http://mathworld.wolfram.com/Chi-SquaredTest.html

[10] Y. Kabutoya, T. Yumoto, S. Oyama, K. Tajima and K. Tanaka, Quality Estimation of Local Contents Based on PageRank Values of Web Pages, In Proceedings of the 22nd ICDEW, pp. 126-129, 2006.

[11] S. Brin and L. Page, The Anatomy of a Large-scale Hypertextual Web Search Engine, In Proceedings of the 7th WWW, pp. 107-117, 1998.