대규모 문서 데이터 집합에서 Q&A를 위한 질의문 분류 기법

엄재홍 장병탁
서울대학교 컴퓨터공학부
{jheom, btzhang}@scai.snu.ac.kr

A Query Classification Method for Question Answering on a Large-Scale Text Data

Jae-Hong Eom Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약
어떤 질문에 대한 구체적 답변을 얻고 싶은 경우, 일반적인 정보 검색이 가능한 문제점은 검색 결과가 사용자가 찾고자 하는 답변을 포함하는 또는 포함하지 않는 문서의 검색이라는 점이다. 사용자가 후보문서를 모두 읽어 볼 필요 없이 빠르게 원하는 정보를 얻기 위해서는 검색의 결과로 문서 검색을 제외하기 보다는 실제 원하는 답변을 제공하는 시스템의 필요성이 대두되었다. 이를 위해 기존의 TF-IDF(Term Frequency-Inversed Document Frequency) 기반의 정보검색의 방식에 자연어처리(Natural Language Processing)을 이용한 질문의 분류와 문서의 사전 표지(Tagging)를 사용할 수 있다. 본 연구에서는 매년 NIST(National Institute of Standards & Technology)와 DARPA(Defense Advanced Research Projects Agency) 주관으로 열리는 TREC(Text Retrieval Conference) 중 1999년에 열린 TREC-8의 사용자 질문(Question)에 대한 답변(Answer)을 찾는 'Question & Answer'문제의 실험 환경에서 질문을 특성별로 분류하고 검색 대상의 문서에 대한 사전 표지를 이용한 정보검색 시스템으로 사용자의 질문(Question)에 대한 답변을 보다 정확하고 효율적으로 제시할 수 있음을 실험을 통해 보인다.

1. 서론

인터넷 등의 발달로 온라인 상에서 검색을 통하여 활용할 수 있는 정보의 양이 많아짐에 따라 정보의 홍수(Information Overload) 문제가 대두되고 있다[1]. 이러한 문제를 대하기 위해서 있는 방법은 검색 엔진을 사용하는 것이다. 이러한 검색 엔진들은 자세적으로 웹(Web)에 있는 자료들에 대한 인덱스를 구축하고 나중에 사용자가 입력하는 질의에 가장 가까운 문서들을 인덱스를 이용하여 검색하여 제공한다. 하지만 검색의 결과로 제공되는 문서의 수가 너무 많고 또 원하는 정보를 위해서 문서들을 모두 읽어보아야 한다는 점에서 원하는 정보를 얻기 위해 많은 시간이 요구되는 단점은 가깝다. 특히, 사용자가 특정한 문제(질문)에 대한 답변만 원하는 경우 검색을 통해 주어지는 문서의 집합은 사용자에게 큰 도움을 주기 어렵다. 이에 따라, 정보를 좀더 명확하고 요약하여 사용자에게 빠르게 임의 추출하여 사용자가 원하는 답변을 제공하기 위한 여러 가지 시도들이 있어왔다. 주어진 질문에 대한 대답을 제시하는 것은 아니지만 자연어처리분야의 문서 자동요약시스템은 사용자에 하급 문서 전체를 읽는 수고를 하지 않아도 되도록 돕는다[2]. 자동요약시스템 `SUMMAR-IST'는 여러 분야의 문서자료(corpus)를 기반으로 요약할 문서의 주제를 선정하여 해당 분야의 사전 어휘를 사용하여 요약 문장을 처리해 준다[3].


본 논문에서는 질문의 특성별 분류와 대상 문서에 대한 최소한의 전처리를 통해서 효율적으로 사용자가 입력한 질문에 대한 답변을 찾아 제시할 수 있음을 실험을
협을 통해 보인다.

2. TREC Q&A

TREC Q&A 문제는 지정된 대규모 문서 데이터 집합에서 입력으로 주어진 질문(Question)에 대해 질문과 관련된, 문서의 리스트가 아닌 대답 해답을 도출하는 것을 목적으로 한다.

2.1 TREC Q&A 데이터

본 실험에서는 NIST와 DARPA 주관으로 열리는 본문검색에 관한 컨퍼런스인 TREC(Text Retrieval Conferences)에서 제공하는 자료(전체 CD 5장) 중 Q&A 문제에 대해 공식적으로 지정된 문서 집합인 4-5번 CD에 있는 자료를 검색 대상으로 실시하였다.

실험에 사용된 TREC 데이터는 표1과 같이 전체 528155개의 영문 문서 집합으로 구성되어 있으며, 각 문서의 본문에는 문서번호나 제목 또는 특정한 문서의 필드(field)나 문서의 구조, 기호들을 나타내기 위해서 SGML(Standard Generalized Markup Language)로 된 구분기호(Tag)가 추가되어 있다.

표 1. 사용된 데이터의 범주별 분류

<table>
<thead>
<tr>
<th>구분</th>
<th>범주</th>
<th>문서 개수</th>
<th>문서내용</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISK 5</td>
<td>FBIS</td>
<td>130471</td>
<td>Foreign Broadcast Information Service (1996)</td>
</tr>
<tr>
<td></td>
<td>LATIMES</td>
<td>131896</td>
<td>Los Angeles Times (1989, 1990)</td>
</tr>
</tbody>
</table>

사용한 문서 전체의 크기는 1.96GB이며, CD 4에는 본래 CR93(Congressional Record of the 103rd Congress 1993)의 자료가 추가적으로 포함되어 있지만 Q&A 문제에서는 사용하지 않았다.

2.2 TREC Q&A 결과계층

정의 문장 완성한 문장으로 주어지며 질의 문에 대한 답은 그 크기를 50바이트(byte)나 250바이트로 하는 두 가지 방식 중 하나로 제시하게 된다. 제시되는 답은 줄쳐 문서의 번호와 함께 5개까지의 후보를 포함한다.

그림 1. 질문의 형식과 답변의 예

3. 효과적인 Q&A를 위한 질의 분류 기법

3.1 질문 분류 (Question Classification)

질문 분류는 크게 16가지로 하였다. 분류된 질문들을 각 질문의 종류에 해당하는 질문 분류표(Q-Tag)를 붙여 본문 검색을 할 때 사용할 수 있도록 하였으며 분류별 질문표기는 표2와 같다.

<table>
<thead>
<tr>
<th>Q-Tag</th>
<th>질문의 유형</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q-NAME</td>
<td>Who/What/Which ~</td>
</tr>
<tr>
<td>Q-COST</td>
<td>How much ~</td>
</tr>
<tr>
<td>Q-LOCATION</td>
<td>Where/What ~</td>
</tr>
<tr>
<td>Q-TIME</td>
<td>When ~</td>
</tr>
<tr>
<td>Q-YES-NO</td>
<td>Is it ~</td>
</tr>
<tr>
<td>Q-DURATION</td>
<td>How long ~</td>
</tr>
<tr>
<td>Q-PERSON</td>
<td>Who ~</td>
</tr>
<tr>
<td>Q-DATE</td>
<td>What date/When ~</td>
</tr>
<tr>
<td>Q-AREA</td>
<td>How big ~</td>
</tr>
<tr>
<td>Q-LENGTH</td>
<td>How long/high/etc ~</td>
</tr>
<tr>
<td>Q-ORGANIZATION</td>
<td>Who ~/What ~</td>
</tr>
<tr>
<td>Q-AGE</td>
<td>How old ~</td>
</tr>
<tr>
<td>Q-WEIGHT</td>
<td>How heavy/big ~</td>
</tr>
<tr>
<td>Q-NUMBER</td>
<td>How many ~</td>
</tr>
<tr>
<td>Q-QUANTITY</td>
<td>How big/much ~</td>
</tr>
<tr>
<td>Q-YEAR</td>
<td>When/What year ~</td>
</tr>
</tbody>
</table>

입력으로 들어오는 200개의 질의문들은 각 질의문의 수행에 따라 파싱(parsing)을 거쳐 표2의 분류로 나뉘었으며, 다음과 시스템에서 [4]와 유사한 후처리를 거쳐 적절한 형태로 변경되었다. 질문에 포함되어있던 다른 단어(term)들에 대해서는 이후 그 단어의 고유성 등에 따라 가중치(weight)가 따로 계산되었다.

3.2 문서 전처리(Target Document Preprocessing)

문서 본문들은 본문의 특정 부분이 표2에서 나온 분류 중에서 어떤 분류에 속하는지에 대해 사전에 표지를 (Tagging) 처리를 하였다. 이들은 질문에 대한 검색을 처리할 때 해당 분류의 질문이 가질 수 있는 단어에 대한 단어의 기존의 해답으로 쓰이게 된다.

3.3 제시된 결과의 평가(Answer Evaluation)

제시된 결과는 각각의 질문에 대해 후보 및 순위에서 정답이 제시되었는지를 기준으로 점수를 부여하였다.

\[
Score(i) = \frac{1}{Rank(i)}
\]

따라서 순위에 따라 각 해당 후보는 1.0, 0.5, 0.33, 0.25, 0.20, 0.0의 점수를 가지게 된다. 평가는 NIST에서 제공한 TREC-8 Q&A 정답 파일을 사용하였다.
4. 실험 및 결과

실험은 TREC-8에서 사용한 원전문 기존 학습에 198개 전체 200개 질문 중 2개를 제외) 이해에 사용한 질문 검색에 있는 2 개를 제외)의 질문을 사용하였다. 실험에 사용된 2851개의 문서에 대한 사전표준화는 초기에 한 번만 수행하였고, 질문에 대한 파생 및 문맥은 매번 동일하며 추가적인 변화 요인이 없으므로 실험은 1번만 수행하였다.

4.1 실험 결과

실험은 동일한 알고리즘을 이용하여 정답거리 250바이트와 50바이트의 두 가지 모두에 대비하였다.

<table>
<thead>
<tr>
<th>표 3. tf-idf 만을 사용한 단순 정답검색</th>
<th>50 byte run</th>
<th>250 byte run</th>
</tr>
</thead>
<tbody>
<tr>
<td>답을 찾은 후보 순위</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>후보 1로 답을 찾은 문서</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>후보 2로 답을 찾은 문서</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>후보 3로 답을 찾은 문서</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>후보 4로 답을 찾은 문서</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>답을 찾지 못한 문서 수</td>
<td>161</td>
<td>154</td>
</tr>
<tr>
<td>평 균</td>
<td>0.105</td>
<td>0.121</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>표 4. 질의문 분류와 문맥표준을 사용한 정답검색</th>
<th>50 byte run</th>
<th>250 byte run</th>
</tr>
</thead>
<tbody>
<tr>
<td>답을 찾은 후보 순위</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>후보 1로 답을 찾은 문서</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td>후보 2로 답을 찾은 문서</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>후보 3로 답을 찾은 문서</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>후보 4로 답을 찾은 문서</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>답을 찾지 못한 문서 수</td>
<td>110</td>
<td>98</td>
</tr>
<tr>
<td>평 균</td>
<td>0.265</td>
<td>0.283</td>
</tr>
</tbody>
</table>

5. 결론

본 논문에서는 사용자의 질의에 대해 여러 판정 대규모 문서 데이터 집합에서 각 질의의 답 찾아서 제시하는 Question Answering 시스템에서, 보다 효율적으로 질의에 대한 답을 검색하기 위한 방향으로 질의문 분류에 대한 시험표준화 및 테스트 문소화 방법을 제시하였다. 또한 TREC-8 환경에서의 실험을 통해서 제시된 방법을 적용한 시스템이 그렇지 않은 시스템보다 정확하게 사용자의 질의에 대한 답을 제시하는 것을 보였다.

표 3, 4에서 볼 수 있듯이 질의문에 대한 분류를 고려하는 경우가 그렇지 않은 경우보다 확실한 성능의 향상 을 보이고 있다. 이러한 결과는 단순히 문서상의 단어의 출현빈도(tf-idf)만 고려한 검색방식보다 질의와 검색 대상 문론을 꼭히 분석하는 방법이 질의에 대한 답을 찾아내는데 보다 효과적이라는 것을 나타낸다고 할 수 있다.

그렇지만 질의문별 휴대 분포(본 논문에서는 분포도를 제시하지 않았을)를 고려하면 중간간간에 수가 상 대적으로 매우 낮은 부분이 여러 곳에서 발견되는데, 이는 아직도 질의문의 분류방식이 최적화되지 못했음을 의미하는 것으로, 이러한 문제의 해결을 위해서는 보다 자세하고 포괄적인 질의분류에 관한 연구가 필요하다. 또한 수동으로 질의를 분류하는 것은 많은 수의 질의문에 대해서는 불가능하므로 기계학습이나 기타 방식의 응용을 통한 질의문 자동 분류에 관한 연구도 더 필요하다.

감사의 글
본 연구는 정보통신부 대학기초 연구(과제번호 98-199)에 의해 일부 지원되었음.

참고문헌