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

어떠한 질문에 대한 구체적 해답을 얻고 싶은 경우, 일반적인 정보 검색이 가지는 문제점은 검색 결과가 사용자가 찾고자 하는 답이라 하기 보다는 해답을 포함하는(또는 포함하지 않는) 문서의 집합이라는 점이다. 사용자가 후보문서를 모두 읽을 필요 없이 빠르게 원하는 정보를 얻기 위해서는 검색의 결과로 문서집합을 제시하기 보다는 실제 원하는 답을 제공하는 시스템의 필요성이 대두된다. 이를 위해 기존의 TF-IDF(Term Frequency-Inversed Document Frequency) 기반의 정보검색의 방식에 자연언어처리(Natural Language Processing)를 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 사용자의 질문(Question)에 대한 답(Answer)을 찾는 'Question & Answer'문제의 실험 환경에서 질문을 특징별로 분류하고 검색 대상의 문서에 대한 사전 표지를 이용한 정보검색 시스템으로 사용자의 질문(Question)에 대한 해답을 보다 정확하고 효율적으로 제시할 수 있음을 실험을 통하여 보인다.

1. 서론

인터넷 등의 발달로 온라인 상에서 검색을 통하여 활용할 수 있는 정보의 양이 많아짐에 따라 정보의 홍수(Information Overload) 문제가 대두되고있다[1]. 이렇게 점차 비대해지고 있는 인터넷에서 사용자가 필요한 정보를 찾기 위해서 사용하는 가장 보편적인 방법은 검색엔진을 사용하는 것이다.

이러한 검색 엔진들은 자체적으로 웹(Web)에 있는 자료들에 대한 인덱스를 구축하고 나중에 사용자가 입력하는 질의에 가장 가까운 문서들을 이 인덱스를 이용하여 검색하여 제공한다. 하지만 검색의 결과로 제공되는 문서의 수가 너무 많고 또 원하는 정보를 위해서 문서들을 모두 읽어(살펴)봐야 한다는 점에서 원하는 정보만을 골라보기에는 너무 많은 시간을 요구한다는 단점을 가진다. 특히, 사용자가 특정한 문제(질문)에 대한 해답만을 원하는 경우 검색을 통해 주어지는 문서의 집합들은 사용자에게 큰 도움을 주기가 어렵다.

이에 따라, 정보를 좀더 압축하고 요약하여 사용자에게 빠르게 읽기 쉬운 형식으로 제공하기위한 여러 가지 시도들이 있어왔다. 주어진 질문에 대한 대답을 제시하는 것은 아니지만 자연언어처리분야의 문서 자동요약시스템은 사용자로 하여금 문서 전체를 읽는 수고를 하지 않아도 되도록 해준다. [2]의 자동요약시스템 'SUMMARIST'는 여러 분야의 문서자료(corpus)를 기반으로 요약할 문서의 주제를 선정하여 해당 분야의 사전 어휘를 사용하여 문서 요약을 처리해 준다[3].

현재까지는 사용자의 질문에 대해서 직접 답을 출력해 주는 시스템에 대한 구체적이고 집중된 연구는 이루어지고 있지 않지만 특정 분야에 대한 FAQ등과 같이 제한된 부분에서 이를 자동화 하기위한 시스템에 대한 연구는 종종 있어왔다. 이러한 시스템의 한 예로 [8]에서 다루고있는 ' FAQFINDER' 라는 시스템을 들 수 있다. 이 시스템은 WordNet[9]을 이용한 질의 확장을 통해서 자동차 분야에 관련된 사용자들의 질문에 대해 기존의 답변 파일 중에서 입력된 질문의 답변으로 가능한 것들을 찾아서 그 해답 문서의 목록을 제시해 주는데, 각 해답 문서는 질문에 대한 대답 형식으로 간결하게 구성되어 있다. [10]의 시스템은 WordNet과 함께 대용량 지식기반(knowledge-base)에서 추론엔진을 이용하여 질문에 대한 대답을 추출해 낸다.

이와 같이 질의에 대해 해답을 출력하는 시스템에 관한 기존 연구들은 주로 추가적인 지식기반에 바탕을 두거나 질의를 이용한 추론을 제어함으로써 그 성능의 향상을 도모 해왔다[8][9][10].

본 논문에서는 질문의 특성별 분류와 대상 문서에 대한 최소한의 전처리를 통해서 효율적으로 사용자가 입력한 질문에 대한서 답을 찾아 제시할 수 있음을 실

험을 통해 보인다.

2. TREC Q&A

TREC Q&A 문제는 지정된 대규모 문서데이터 집합에서 입력으로 주어지는 질문(Question)에 대해 질문과 연관된, 문서의 리스트가 아닌 대한 해답을 도출하는 것을 목적으로 한다.

2.1 TREC Q&A 데이터

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

실험에 사용된 TREC 데이터는 표1과 같이 전체 528155개의 영어 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 추가되어있다.

표 1. 사용된 데이터의 범주별 분류

구분	범주	문서 개수	문서내용
DISK 4	FR 94	55631	Federal Register (1994).
	FT 91~94	210158	Financial Times (1991 ~ 1994 일부)
DISK 5	FBIS	130471	Foreign Broadcast Information Service (1996)
	LATIMES	131896	Los Angeles Times (1989, 1990)

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 결과제출

질의 문은 완전한 문장으로 주어지며 질의 문에 대한 답은 그 크기를 50바이트(byte)나 250바이트로 하는 두 가지 방식 중의 하나로 제시하게 된다. 제시되는 답은 출처 문서의 번호와 함께 5개까지의 후보를 포함한다.

그림 1. 질문의 형식과 답변의 예

```
<top>
<num> Number: 1
<desc> Description:
Who is the author of the book, "The Iron Lady: A Biography of
Margaret Thatcher"?
</top>
```

```
Answer of Question 1:
LA111289-0002 1 210677 ...THE IRON LADY: A BIOGRAPHY
OF MARGARET THATCHER BY HUGO YOUNG (...
FT943-8867 2 200395 ...dubbed "the Iron Lady" the book
of HUGO YOUNG was a...
```

3. 효과적인 Q&A를 위한 질의 분류 기법

3.1 질의문 분류 (Question Classification)

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

표 2. 질문의 유형별 분류

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

입력으로 들어오는 200개의 질의문 들은 각 질문의 유형에 따라 파싱(parsing)을 거쳐 표2의 분류로 나뉘어지며, 다음으로 시스템에서 [4]와 유사한 후처리를 거쳐 적절한 형태로 변경되어진다. 질의에 포함되어있던 다른 단어(term)들에 대해서는 이후 그 단어의 고유성 등에 따라 가중치(weight)가 따로 계산되어진다.

3.2 문서 전처리(Target Document Preprocessing)

문서 본문들은 본문의 특정 부분이 표2에서 나눈 분류 중에서 어떤 분류에 속하는지에 대해 사전에 표지(Tagging) 처리를 하였다. 이들은 질의문에 대한 검색을 처리할 때 해당 분류의 질문이 가질 수 있는 답인지에 대한 판단의 기준의 하나로 쓰이게 된다.

3.3 제시된 결과의 평가(Answer Evaluation)

출력된 결과는 각각의 질문에 대해 후보 몇 순위에서 정답이 제시되었는지를 기준으로 점수를 부여하였다.

$$Score(i) = \frac{1}{Rank(i)}$$

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

4. 실험 및 결과

실험은 TREC-8에서 사용한 완전한 문장으로 이루어진 198개(전체 200개 질문 중 모호성이 있는 2개를 제외)의 질의문을 사용하였다. 실험에 사용된 528155개의 문서에 대한 사전표지는 초기에 한번만 수행하였고, 질의문에 대한 파싱과 분류는 매회 동일하며 추가적인 변화요인이 없으므로 실험은 1번만 수행하였다.

4.1 실험 결과

실험은 동일 알고리즘을 이용하여 정답크기 250바이트와 50바이트의 두 가지 모두에 대해 실시하였다.

표 3. tf-idf 만을 사용한 단순 정답검색

답을 찾은 후보 순위	50 byte run	250 byte run
후보 1로 답을 찾은 문제	11	13
후보 2로 답을 찾은 문제	12	12
후보 3로 답을 찾은 문제	4	5
후보 4로 답을 찾은 문제	8	11
후보 5로 답을 찾은 문제	2	3
답을 찾지 못한 문제 수	161	154
평 균	0.105	0.121

표 4. 질의문 분류와 본문표지를 사용한 정답검색

답을 찾은 후보 순위	50 byte run	250 byte run
후보 1로 답을 찾은 문제	33	33
후보 2로 답을 찾은 문제	21	24
후보 3로 답을 찾은 문제	12	12
후보 4로 답을 찾은 문제	13	18
후보 5로 답을 찾은 문제	9	13
답을 찾지 못한 문제 수	110	98
평 균	0.265	0.283

5. 결론

본 논문에서는 사용자의 질의에 대해 미리 주어진 대규모 문서 데이터 집합에서 각 질의의 답을 찾아서 제시하는 Question Answering 시스템에서, 보다 효율적으로 질의에 대한 답을 검색하기 위한 방법으로 검색 본문에 대한 사전표지와 특성별 질의문 분류 방법을 제시하였다. 또한 TREC-8 환경에서의 실험을 통해서 제시된 방법을 적용한 시스템이 그렇지 않은 시스템보다 정확하게 사용자의 질의에 대한 답을 제시하는 것을 보았다.

표3,4에서 볼 수 있듯이 질의문에 대한 분류를 고려하는 경우가 그렇지 않은 경우보다 확실한 성능의 향상을 보이고있다. 이러한 결과는 단순히 문서상의 단어의 출현빈도(tf-idf)를 고려한 검색방식보다 질의와 검색대상 본문을 좀더 분석하는 방법이 질의에 대한 답을 찾아내는데 보다 효과적이라는 것을 나타낸다고 할 수 있다.

그렇지만 질의문별 해답 분포(본 논문에서는 분포도를 제시하지 않았음)를 고려하면 중간중간에 점수가 상대적으로 매우 낮은 부분이 여러 곳에서 발견되는데, 이는 아직은 질의문의 분류방식이 최적화되지 못했음을 의미하는 것으로, 이러한 문제의 해결을 위해서는 보다 자세하고 포괄적인 질의 분류에 관한 연구가 필요하다. 또한 수동으로 질의를 분류하는 것은 많은 수의 질의문에 대해서는 불가능하므로 기계학습이나 기타 방식의 응용을 통한 질의문 자동 분류에 관한 연구도 더불어 필요하다.

감사의 글

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

참고문헌

[1] Pattie Maes, " Agents that reduce work and information overload ", Communications of the ACM Vol. 37, No 7, pp.30-31, 1994.

[2] Inderjeet Mani, Mark T. Maybury, *ADVANCES IN AUTOMATIC TEXT SUMMARIZATION* , summarization, 1999.

[3] Eduard Hovy, Chin -Yew Lin, " Automated Text Summarization in SUMMARIST " , in *Proceedings of the Workshop of Intelligent Scalable Text Summarization*, July 1997.

[4] Chin-Yew Lin, Eduard Hovy, " Identify Topics by Position " , in *Proceedings of the 5th Conference on Applied Natural Language Processing*, March 1997.

[5] D. -H. Shin, B. -T. Zhang, "A Two -Stage Retrieval Model for the TREC -7 Ad Hoc Task", *Draft of Text Retrieval Conference -7*, pp. 501-???, November 1998.

[6] J. Prager, D. Radev, E. Brown, A. Coden, "The Use of Predictive Annotation for Question Answering in TREC8", in *Proceedings of Text Retrieval Conference-8*, pp. 323-335, 1999.

[7] D-H Shin, Y-H Kim, S. Kim, J -H Eom, H-J Shin, B-T Zhang, "SCAI TREC-8 Experiments", *Draft of TREC-8*, page ???-???, 1999.

[8] Kristian Hammond, Vladimir Kulyukin, Steven Lytinen, Noriko Tomuro, and Scott Schoenberg, "Question Answering from Frequently -Asked Question Files: Experiences with the FAQ Finder System", *Univ. of Chicago, Department of Computer Science Technical Report TR -97-05*, 1997.

[9] Miller, G. A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* , 38(11).

[10] Sanda Harabagiu and Dan Moldovan, "An Intelligent System for Question Answering ", in *the Proceedings of the 5th Conference on Intelligent Systems*, June 1996, Reno NV, pages 71-75.