韓國 컴퓨터情報學會 論文誌
第10卷 第3號, 2005. 7.

2005-10-3-1-1

개념 속성 기반 정보 검색

윤 보 현*, 서 창 호**

Concept and Attribute based Answer Retrieval

Bo-Hyun Yun* Chang-ho Seo**

요 약

본 연구에서는 지식검색을 위해 개념 속성을 이용하여 사용자 질의에 가장 적합한 정답 문장들을 검색할 수 있는 정답검색 시스템을 설계하고 평가한다. 이 시스템은 먼저 사용자 질의를 개념 속성에 대한 불리언 연산으로 분석한 다음, 정답 문서 색인 집합에서 해당 문서들을 검색한다. 사용자는 이 검색된 문서들로부터 자신이 요구한 정답 문장들을 검색할 수 있으며, 또한 특정한 문서를 선택함으로써 그 문서에 포함된 정답 문장들을 검색할 수 있다. 이를 위해서 개념어와 속성어의 색인 단위로 색인된 정답 문서들은 각각의 문장들로 분할되어 색인된다. 그래서 분할된 문장들은 개념어와 속성어 형태로 분석되어 문서 색인 단위와의 관련 정도를 평가함으로써 정답 문장들의 위치를 색인한다. 마지막으로, 100개의 사용자 질의에 대해 정답 검색 시스템의 성능을 다양한 방법으로 평가한다.

Abstract

This paper presents the information retrieval system which can retrieve the most appropriate answer sentence for user queries by using the concept and the attribute for the knowledge retrieval. The system analyzes the user query into the Boolean queries with the concept and the attribute and then retrieve the relevant documents in the indexing set of answer documents. Users can retrieve the relevant answer sentences from the relevant documents. For this, the answer documents indexed by the concept and the attribute are segmented by each sentence respectively. Thus, the segmented sentences are analyzed into the concept and the attribute of which the relevance degree with indexing units of documents is evaluated. Then, the system indexes the location of answer sentences. In the experiment, we evaluate the performance of our answer retrieval system against 100 user queries and show the experimental results.

▸ Keyword : Information Retrieval, Indexing, Concept, Relevance Degree

---

• 제1저자 : 윤보현
• 접수일 : 2005.03.28, 심사완료일 : 2005.05.21
* 목원대학교 컴퓨터교육과 조교수, ** 공주대학교 바이오정보학과 부교수
※ 이 논문은 2003년도 한국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2003-003-D00378).

Ⅰ. 서 론

현재의 정보검색 시스템은 사용자가 제시하는 질의어 또는 논리 연산자와 결합된 단어집합을 입력으로 하여 정답이 포함되어 있을 가능성이 높은 문서를 제시하고 사용자가 제시된 문서 내에 정답이 포함되어 있는지의 여부를 판단하도록 하고 있다. 그에 반해 질의응답 시스템의 경우는 일반 사용자들에게 익숙한 자연어 문장 형태의 질의를 받아, 정답 또는 정답이 포함된 문장 등을 제시한다. 따라서 사용자에게 보다 높은 정확한 정보를 제공한다는 점에서 그 요구가 증가하고 있다.

본 연구의 목표는 정보검색 시스템 보다는 많은 언어분석과 자원이 필요하지만 문서가 아닌 정확한 정답을 찾을 수 있고, 질의응답 시스템에 비해서 개체명인식 등의 복잡한 언어분석과 자원을 필요로 하지 않으면서 질의응답시스템과 비슷한 정확도를 보이는 시스템을 개발하는 것이다. 정보검색에서 데이터베이스에 들어 있는 검색 결과들은 그 웹 문서의 실질적인 내용보다는 특정 단어만을 검색하여 구축한 결과이므로 사용자가 검색하여 얻은 웹 문서의 신뢰도가 낮아진다. 질의응답에서 원하는 결과를 얻기 위해서는 대용량의 지식베이스를 구축하여야 하며, 자연어분석도 형태소분석, 품사태깅, 개체명인식 혹은 의미분석 등의 복잡한 언어분석을 수행하여야만 원하는 정답을 얻을 수 있다.

이와 같은 기존 연구의 문제점을 바탕으로 본 논문에서는 개념 속성 기반 정답 검색 방안을 제안한다. 개념 속성을 이용한 정답 검색은 특정한 개념에 대해 사용자들이 실제 알고 싶은 정보들을 그 개념의 속성에 따라 분류된 정답 문서 집합으로부터 정답을 검색하는 것으로 정의한다. 일반적으로 사용자들은 질의에 표현되는 몇 개의 개념에 대한 포괄적인 검색을 요구하는 것이 아니라 그 개념들이 가지는 특정한 속성으로 분류되는 부분적인 정보만을 요구한다. 따라서, 이 정보들만을 선택적으로 검색할 수 있도록 지원하는 검색 방법은 인터넷이나 도서관과 같은 방대한 문서 검색에서 반드시 요구되는 기능이다. 또한, 사용자들이 검색된 문서의 전체 내용보다는 특정 부분만을 요구하는 경우가 빈번하기 때문에 하나의 검색된 문서에서 사용자의 요구에 정확히 일치하는 부분만을 제시할 수 있는 기능 역시 요구된다.

본 논문에서는 정답 문서 집합은 특정 개념과 관련된 모든 문서들 중에서 개념의 속성에 따라 분류된 문서들로 정의하며, 하나의 문서에서 사용자 질의와 가장 일치하는 부분을 정답 위치로 각각 정의한다. 사용자 질의와 가장 일치하는 정답 문서를 검색하는 방법과 정답 위치를 추출하는 방법에 대해 설명한다.

Ⅱ. 관련연구

본 연구의 관련연구로서 정보검색시스템과 질의응답시스템에 대해서 알아본다. 정보검색 시스템은 일반적으로 검색어라는 일련의 단어들을 이용하여 그 단어가 들어있는 웹 문서들을 검색해준다[1, 3, 4, 15, 16]. 일반적으로 검색어로 지정해준 단어들을 검색할 때에는 웹 문서의 본문에서 찾지만, 제목(title)이나 웹주소(URL)에서 특정 단어를 검색할 수도 있다[2, 5, 9]. 웹 인덱스방식에서 정보를 찾는 작업은 사용자가 검색어를 입력했을 때에 수행되는 것이 아니다. 이러한 방식을 사용한다면 너무나 많은 시간과 노력이 필요하게 되어서 비효율적이기 때문이다. 웹 인덱스방식은 로봇을 이용하기 때문에 인덱스 데이터베이스가 지나치게 커지게 되고, 검색 결과가 너무 많아져서 오히려 사용자의 판단을 흐리게 할 수도 있다. 또한 데이터베이스에 들어 있는 검색 결과들은 그 웹 문서의 실질적인 내용보다는 특정 단어만을 검색하여 구축한 결과이므로 사용자가 검색하여 얻은 웹 문서의 신뢰도가 낮아진다.

아울러 색인이 정확하지 않은 검색엔진의 경우 원하는 정보를 찾을 수 없는 경우도 있다. 때로는 너무나 많은 검색결과들을 보여줌으로써 실제로 원하지 않는 결과도 함께 보여 주기 때문에 키워드 선정과 연산자 사용이 적절해야한다. 그런데 검색시에 사용하는 연산자들이 초보자들에게는 상당히 어렵기 때문에 단점으로 지적되곤 한다. 색인의 정확도를 높이기 위해 자연어 분석을 통해 구 단위의 색인어를 추출하는 연구도 진행중이지만 실용적으로 사용하기에는 색인어 추출 시간이 많이 소요된다[7, 12, 13, 19, 21].

질의응답 시스템은 크게 질의문 분석과 정답추출이 시스템의 핵심을 이루고 있다[14, 17, 18, 20]. 기존 연구 중 질의문 분석에 관하여서는 의문사 등의 키워드 추출 및 가중치 부여와 개체인식 기법에 관한 연구가 이루어져왔으며,

정답추출 부분에서는 고유명사에 대한 인식과 추정명사 등에 대한 의미속성 결정 등을 위한 대규모 지식베이스 구축 및 활용방법과 함께 정확한 정답추출을 위한 구문분석에 관한 활발한 연구가 이루어지고 있다. 영어권에서는 워드넷을 이용한 지식베이스 관련연구가 이루어지고 있으나 국내에서는 워드넷과 같은 언어자원의 부족으로 영어권에 비해 낮은 수준의 실험결과를 보이고 있다[6, 8, 10, 11].

III. 개념 기반 정보 검색 방안

다음은 시스템 구성도를 통해 이들의 주요 구성과 상호 관계를 설명한다.

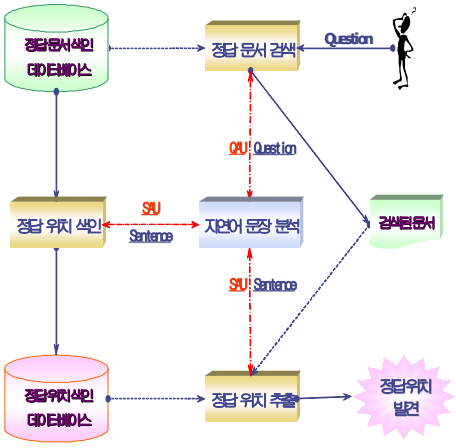


그림 1. 정보 검색 시스템 구성도
Fig 1. Information Retrieval System Configuration

(그림 1)은 정답 문서들을 검색하고 정답 위치를 추출할 수 있는 정답 검색 시스템의 구성을 나타내고 있다. 특히, "정답 문서 검색 엔진", "정답 위치 색인 엔진" 그리고 "정답 위치 추출 엔진" 사이의 관계를 설명하고 있으며, 이들은 하나의 문장에서 개념 용어와 속성 용어들을 추출할 수 있도록 지원하는 "자연어 문장 분석" 모듈을 기본적으로 이용하게 된다.

"정답 문서 색인 데이터베이스"는 앞 절에서 설명된 분류 방법을 통해 모든 문서들을 "[개념 용어, 속성 용어] → [DocID]" 형태로 색인한 데이터베이스이며, 정답 위치 색인 데이터베이스는 하나의 문서에 포함된 문장들이 문서 색인 단위 [개념 용어, 속성 용어]와 어느 정도 관련이 있는지를 "[개념 용어, 속성 용어] → [DocID][SentNum,Weight]" 형태로 색인된 데이터베이스이다. 여기서, 정답 위치 색인 엔진에 의해 색인되는 문장 SentNum의 "Weight"는 문서 색인 단위 [개념 용어, 속성 용어]와 문서가 포함하는 문장들의 [개념 용어, 속성 용어]에 대한 관련 정도를 나타낸다[4].

정답 문서 검색 엔진은 사용자 질의를 [개념 용어, 속성 용어, 보조 용어]들의 불리언 형태로 분석한 결과와 개념적으로 관련이 있는 문서 색인 단위로 색인된 문서들을 문서 색인 데이터베이스로부터 검색한다. 이 검색된 문서들 중에서 사용자가 특정 문서를 선택하면, 정답 위치 추출 엔진이 정답 위치 색인 데이터베이스에서 "Weight > 임계값"인 문장들을 분석한 결과와 질의 분석 결과 사이의 관련 정도를 평가한다. 이 값은 문장과 질의 사이의 관련 정도를 나타내며, 가장 높은 문장을 정답 문장으로 제시하게 된다.

3.1 자연어 문장 분석

자연어 문장 분석 모듈은 사용자 질의나 문서에서 추출된 하나의 문장으로부터 아래와 같은 특정한 형태의 분석 정보를 추출하게 된다. 이 모듈의 두 가지 형태의 분석 결과 QAU(Question Analysis Units)과 SAU(Sentence Analysis Units)은 정답 문서 검색과 정답 위치 추출에서 사용자 질의 처리와 문서에 대한 문장 분석을 위해 각각 이용된다.

$$AU_i = \begin{matrix} \text{analysis\_point} : [\text{keyword}, \text{attrword}] \\ \text{auxiliary\_point aux\_word} \end{matrix}$$

$$QAU = AND|OR_{i=0}^{n} AU_i$$

$$SAU = \sum_{i=0}^{n} AU_i$$

여기서, QAU는 질의를 분석한 결과이며, SAU는 문서를 구성하는 자연어 문장을 분석한 결과이다. 이들의 자세한 구성은 다음과 같이 표현할 수 있으며, 이들은 정답 위치 색인, 정답 문장 검색 그리고 정답 위치 추출에 중요한 모듈로 이용된다.

여기서, 질의 분석 결과인 QAU는 AU들에 대한 논리곱 또는 논리합의 조합으로 구성되며, SAU는 정답 문장 색인 정보로 이용되기 때문에 단지 AU의 연속으로 표현된다. 또한, 하나의 AU는 개념 용어(keyword)와 속성 용어(attrword)의 쌍으로 구성되는 핵심 관점(analysis_point) 그리고 핵심 용어로는 사용되지 않지만 하나의 문장에 대한 보조 정보를 제공하는 용어(aux_word)들로 구성되는 보조 관점(auxiliary_point)으로 표현된다. 따라서, QAU나 SAU 분석하기 위한 문장은 모두 병렬 구조를 가진다고 가정하였다[15].

SAU의 분석 방법은 QAU 보다 단순하기 때문에 다음은 QAU 분석 방법에 대해 예를 통해 설명한다. 먼저, 논리적 연산자를 기준으로 하나의 문장을 AU 단위로 분할한다. 이를 구분하기 위해 "~이며", "~이고", "~과", "그리고", "또는", "이거나" 등과 같은 형태소 분석의 성분들을 이용한다. 이 성분들은 QAU의 논리곱이나 논리합으로 변형될 수 있다. 다음으로 각각의 AU 단위에는 개념 용어와 속성 용어가 각각 하나씩 존재하게 되며 이들은 개념 용어는 개념 용어 사전으로부터 식별되며, 속성 용어는 속성 용어 사전과 속성 패턴 사전으로부터 식별된다. 예를 들어 하나의 질의 Q에 대해 다음과 같이 분석될 수 있다.

Q = 투자신탁회사는 무엇**이며** 높은 수익의 투자신탁에는 어떤 상품이 있는가?
→ [**투자신탁회사**는 **무엇이며**] [높은 수익의 **투자신탁**에는 **어떤 상품이 있는가**]?

여기서, "~이며"는 AND로 분류되고 투자신탁회사와 투자신탁은 각각 개념 용어로 설정된다. 또한, "무엇이며"와 "어떤 상품이 있는가"는 속성 영역으로서 개념에 대한 속성 용어를 결정하게 된다. 즉, "무엇이며"는 속성 패턴 데이터베이스로부터 속성 워드 "정의"로 매핑되며, "어떤 상품이 있는가"는 속성 영역에 "상품"이라는 속성이 존재하기 때문에 속성 용어가 결정된다. 또한, "높은"이나 "수익"은 투자신탁의 상품에 대한 주변 정보이기 때문에 주변 용어로 식별된다. 따라서, Q는 AU1=[투자신탁회사, 정의,{null}] 그리고 AU2= [투자신탁, 상품,{높다, 수익}]을 가지며 이 두 AU는 AND 연산자로 연결된다. 따라서 다음과 같은 QAU를 얻을 수 있다.

QAU = [투자신탁회사, 정의,{null}] AND [투자신탁, 상품,{높다, 수익}]

다른 예로 Q="투자신탁회사는 무엇이며 어떤 것을 판매하는가?"에 대해 "~이며"를 중심으로 AU1=[투자신탁회사, 정의,{null}]이고 AU2= [null, 상품, {null}]이다. 또한, "~이며"는 AND로 매핑되기 때문에 QAU=AU1 AND AU2로 분석된다. 이때, AU2는 개념 용어가 null인 경우, 병렬 문장에서는 왼쪽 AU의 개념 용어가 오른쪽 AU의 개념용어로 파생되는 구조적인 특성을 가지기 때문에 AU2=[투자신탁회사, 상품, {null}]로 결정된다. Q="견질어음과 백지어음에 대한 의미 차이는 무엇인가?"와 같은 질의에 대해서는 "~과"를 중심으로 AU1={견질어음, null, {}}이고 AU2={백지어음, 정의, {차이}}이다. 또한, "과"는 AND로 매핑되기 때문에 QAU=AU1 AND AU2로 분석될 수 있다. 이때, AU1의 속성 용어가 null인 경우, 병렬 문장에서는 오른쪽 AU의 속성 용어가 왼쪽 AU의 속성 용어로 파생되는 구조적인 특성을 가지기 때문에 AU2의 속성 용어가 AU1의 속성 용어로 파생된다.

자연어 분석 모듈의 기본 기능을 나타내고 있다. 이 모듈은 형태소 분석기와 같은 자연어 처리 도구를 이용하며, 속성 추출을 위해 각 속성에 해당하는 기본 패턴에 대한 데이터베이스가 요구된다. 이 자연어 문장 분석을 위해 본 연구에서는 12만개의 엔트리를 가지는 개념 용어 사전, 89개의 엔트리를 가진 속성 용어 사전 그리고 각각의 속성에 대한 패턴 사전을 이용하였다. 또한, 속성 용어는 모든 개념에 적용될 수 있는 일반 속성과 특정 개념에만 적용될 수 있는 개별 속성으로 각각 구분된다.

3.2 정답 문서 검색

정답 문서 검색 엔진은 질의 분석 결과의 QAU와 일치하는 정답 문서 색인 데이터베이스의 색인 단위 [개념 용어, 속성 용어]로 색인된 문서들을 사용자에게 제시한다. 이때, 개념 용어들 사이의 의미 관계를 정의한 개념망이 이용되어 개념 용어와 의미적으로 밀접한 관계를 가지는 개념 용어들로 색인된 정답 문서 역시 검색될 수 있다는 특징을 가진다. 다음은 정답 문서 검색 엔진의 기본 구성을 살펴보고 이들을 구성하는 각각의 모듈에 대해 설명한다.

먼저, 사용자 질의는 자연어 분석 모듈을 통해 QAU로 분석되며, 이 QAU를 구성하는 AU로 색인된 정답 문서를 정답 문서 색인 데이터베이스로부터 검색한다. 이때 사용되는 불리언 연산자는 각각 퍼지 합집합과 퍼지 교집합으로 해석된다. 또한, 이 엔진은 개념망을 통한 확장 검색을 수행한다. 즉, 이 검색에서는 AU의 개념어와 하위 또는 형제 관계를 가지는 개념어들을 개념망을 통해 추출한 다음, 이들로 색인된 정답 문서들까지 검색할 수 있다[14].

예를 통해, 정답 문서 검색 과정을 설명하고 있다. 먼저, 사용자 질의 Question ="투자신탁회사는 무엇이며 높은 수익의 투자신탁에는 어떤 상품이 있는가?"는 자연어 문장 분석 모듈을 통해 QAU = [투자신탁회사, 정의,{}] AND [투자신탁, 상품,{높다, 수익}]로 분석된다. 이 분석된 QAU의 각각의 AU는 개념망을 통해 다음과 같이 확장되어 검색에 이용된다. 즉, "투자신탁회사"의 하의어는 "로베코, " 현대투자신탁증권" 등이 있고, "투자신탁"의 하위어는 "뮤추얼 펀드", "벤처펀드" 등이 있기 때문에 QAU는 확장될 수 있다.

다음은 확장된 QAU에 의해 검색된 문서들이며 "1/0.8" 은 1번 문서와 색인 단위 [투자신탁회사, 정의]의 관련 정도를 나타낸다. 또한, ||QAU||에 의해 평가된 "1/0.732" 는 두 색인 단위[투자신탁회사, 정의]과 [투자신탁, 상품]에 의해 동시에 색인된 문서 1과 관련 정도를 나타내며 이 관련 정도는 아래의 p-매칭 함수에 의해 일관되게 계산된다.

||QAU|| = |||[투자신탁회사, 정의] AND [투자신탁, 상품]||| = {**1/0.8**, 2/0.3, 3/0.8,…} ∩ {1/0.7, 3/0.9, 5/0.3,…} = {**1/0.732**, 3/0.842,…}

$$sim_{\cup}(x_1,x_2)=\left(\frac{x_1^{\,p}+x_2^{\,p}}{2}\right)^{p} \qquad sim_{\cap}(x_1,x_2)=1-\left(\frac{(1-x_1)^{p}+(1-x_2)^{p}}{2}\right)^{p}$$

3.3 정답 위치 색인 엔진

정답 위치 색인 엔진은 문서 색인 단위와 각 문장 사이의 의미적인 관련 정도를 평가하여 색인한다. 이때, 문서 색인 단위의 개념 용어와 분석된 문장 SAU의 개념 용어 사이의 관련 정도는 개념망을 통해 평가된다. 정답 위치 색인 데이터베이스는 [개념 용어, 속성 용어]→[문서 번호][문장 번호/관련 정도]로 구성되며, 이 엔진을 통해 정답 위치 색인 데이터베이스의 [문장 번호/관련 정도]가 설정된다. 이 관련 정도는 정답 문서 검색에서 검색된 문서를 순위화하기 위해 사용된다. 즉, 정답 문서 색인 단위와 가장 일치하는 문장의 관련 정도를 이 과정에서 평가할 수 있기 때문에 이들에 따라 검색된 정답 문서를 순위화하여 제시한다. 따라서, 가장 정답이 존재할 가능성이 높은 문서를 사용자에게 먼저 보여줌으로써 다음 장에서 설명될 로컬 정답 문장을 사용자가 쉽게 검색할 수 있도록 지원한다.

먼저, 정답 문서 색인 데이터베이스에서 특정한 문서 색인 단위(IU: Indexing Unit)로 색인된 문서를 추출하여, 문장 단위로 분할한다. 이 분할된 문장들은 자연어 문장 분석 모듈을 통해 SAU로 분석된다. 따라서, SAU Matching 모듈은 이 문서를 색인하는 단위인 IU와 각 문장들의 SAU의 매칭을 통해 일정한 임계값 이상의 문장 번호와 그 관련 정도를 데이터베이스에 색인하게 된다.

문장 분할은 다음과 같은 기본적인 휴리스틱만을 이용하고 있다. 즉, 문장 분리자로 ".", "?", 그리고 "!"를 기본적으로 사용하며, "~다", "~음", "~함" 등과 같은 종결형 어미 역시 이용한다. 또한, 하나의 문장의 최대 700 문자를 초과할 수 없으며, 14 문자 이하가 될 수 없다. 특히, 정확한 문장 분할은 자연어 문장 분석 모듈이 적절한 SAU를 추출하는 기본이 되기 때문에 매우 필요한 모듈이다. 이 기술은 현재 자연어 처리에 대한 많은 연구에서 제시하고 있으며, 향후 가장 적절한 알고리즘의 개발이 필수적이다. 더욱이, WEB 문서를 처리하는 본 기술에는 HTML 문서의 구조적 특성을 역시 이용할 필요가 있다. 다음은 예를 통해 정답 위치 색인 엔진에 대해 설명한다.

정답 위치 색인 과정을 설명하고 있다. 먼저, 정답 문서 색인 데이터베이스에서 색인되지 않은 문서의 색인 단위와 문서 번호를 추출한 다음, 문서를 문장 단위로 분할하고 문서 문장 처리에 의해 이 문장들은 SAU로 분석된다. 이 분석된 SAU는 색인 단위와 매칭을 통해 문장과 색인 단위 사이의 관련 정도를 평가하게 된다. 색인된 문장 번호와 관련 정도는 색인 정보 저장 단계를 통해 정답 위치 색인 데이터베이스에 저장된다.

문장과 색인 단위의 관련 정도는 색인 단위 IU=[keyword, attrword]와 SAU의 AU들에 대한 핵심 관점(analysis_point) 매핑의 최대값으로 결정된다. 즉, 개념 용어(keyword) 매칭은 개념망에서 두 용어 사이의 거리 DIST를 통해 $w=\alpha\times e^{-0.2\times DIST(keyword_1,keyword_2)}$ $(0\le\alpha\le1)$로 결정된다. 또한, w가 임계값 $\delta(0\le\delta\le1)$ 보다 크고 속성 용어가 서로 같을 경우 w의 값을 $1-\alpha$증가시킨다. 예를 들어, AU1은 개념 용어만 일치하기 때문에 IU와 α의 관련 정도를 가지는 반면, AU2는 속성 용어까지 일치하기 때문에 1.0의 관련 정도를 가진다.

특히, $CNet(keyword_1,keyword_2)=e^{p\times DIST(keyword_1,keyword_2)}$ 은 개념망에서 두 용어 사이의 관련 정도를 계산하는 함수이다. (그림 2)에서 p는 가장 완만한 포물선을 그리는 그래프의 값 -0.2로 결정하였다. 여기서, x축은 두 개념 사이의 거리를 나타내며, y축은 개념 거리에 대한 관련 정도를 0과 1사이의 관련 정도로 나타낸 값이다. 이 값은 많은 실험을 통해 가장 적절한 값으로 향후 대치될 수 있다.

3.4 정답 위치 추출 엔진

정답 위치 추출 엔진은 사용자 질의와 가장 유사한 문서의 일부분을 사용자에게 제시함으로써 검색된 문서를 사용자가 전부 읽어보지 않더라도 문서에서 자신이 요구한 부분을 추출할 수 있도록 지원한다. 이를 위해서는 질의와 문서에 포함된 문장들 사이의 관련 정도를 평가할 수 있어야 한다.

	0.1	0.2	0.3	0.4	0.5
0	1	1	1	1	1
1	0.9048	0.8187	0.741	0.67	0.607
2	0.8187	0.6703	0.549	0.449	0.368
3	0.7408	0.5488	0.407	0.301	0.223
4	0.6703	0.4493	0.301	0.202	0.135
5	0.6065	0.3679	0.223	0.135	0.082
6	0.5488	0.3012	0.165	0.091	0.05
7	0.4966	0.2466	0.122	0.061	0.03
8	0.4493	0.2019	0.091	0.041	0.018
9	0.4066	0.1653	0.067	0.027	0.011
10	0.3679	0.1353	0.05	0.018	0.007

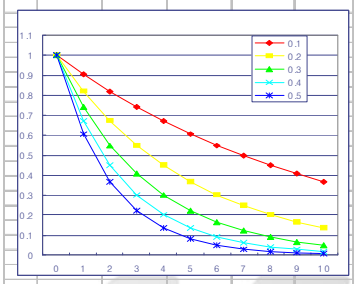


그림 2 개념망에서 두 용어 사이의 관련 정도 평가
Fig 2. Evaluation of Relevance Degree between Two Terms in Concept Network

정답 위치 추출 엔진에 의해 검색되는 정답은 크게 전역 정답과 지역 정답으로 구분된다. 전역 정답은 정답 문서 검색 엔진에서 검색된 모든 문서들로부터 질의와 일치하는 문장들로 추출된 정답들이며, 지역 정답은 특정 문서를 사용자가 선택하였을 경우 그 문서에서 질의와 가장 일치하는 문장들로부터 추출된 정답들이다. 즉, 정답은 질의와 일치하는 문장의 상위 몇 개의 문장과 하위 몇 개의 문장으로 구성된다. 그런데, 이 두 종류의 정답은 결국 지역 정답을 추출하는 방법이 기본으로 이용되기 때문에 여기에서는 지역 정답 추출 방법에 대해 설명한다. 즉, 전역 정답들은 지역 정답들에 대한 퍼지 합집합으로 쉽게 구할 수 있다.

사용자 질의는 QAU로 분석되고, 검색된 정답 문서들 중 하나의 문서에 대해 정답 위치 색인 정보를 처리한다. 이 정답 위치 색인 정보들 중 상위 n개에 대한 정답 문장을 추출하여 분석한다. 이 문장의 분석된 결과인 SAU와 QAU를 매칭하여 가장 일치하는 문장 n개를 추출한다. 이때, QAU와 SAU의 개념어들은 개념망에 의해 매칭된다. 이렇게 매칭된 정답 문장을 기준으로 상하위 문장을 정답으로 제시한다. 전역 정답에 대해서는 이러한 과정을 검색된 정답 문서 각각에 대해 수행하여 가장 높은 관련 정도를 가지는 문장을 기준으로하여 정답을 추출한다.

```
max_w = 0.0;
sent_num = 0;
SAU_LIST = Doc_Sent_Processing(DocID, Θ); ..... (1)
for each SAU in SAU_LIST;
    w = qau_sau_matching(QAU, SAU); ..... (2)
    if(w >= max_w) max_w = w, sent_num = SAU.get_sent_num(); ..... (3)
for end;
return sent_num;
```

```
stack W;
for each q_au in QAU;
        switch for type of q_au;
        case TAU    :  w=au_sau_matching(q_au,SAU); W.push(w); ..... (4)
        case TAND   :  w1=W.pop(); w2=W.pop(); w=sim_AND(w1, w2); W.push(w);
        case TOR    :  w1=W.pop(); w2=W.pop(); w=sim_OR(w1, w2); W.push(w);
for end;
return W.pop();
```

정답 위치 추출을 위한 과정을 예를 통해 설명하고 있다. 사용자 질의는 질의 처리를 통해 QAU로 분석되고, 문서 번호(DocID)를 통해 문서의 문장들을 분할하여 SAU를 분석한다. 이 분석된 QAU와 각각의 SAU의 매칭을 통해 가장 일치하는 정답 문장, 정답 위치 그리고 정답 내용을 제시한다. 정답 위치는 (정답 문장 - pre_n)번째 문장에서 (정답 문장 + post_n)번째 문장으로 결정되며, pre_n과 post_n은 도메인에 특성에 적합한 값을 선정해야 한다. 여기서, 색인 정보 처리는 정답 위치 색인 정보들 중에서 상위 10 문장을 선택함으로써 분석될 문장들을 결정하게 된다. 문서와 질의 사이의 관련 정도를 평가하여 가장 일치하는 문장을 추출하는 알고리즘은 식(1)-(5)과 같다.

여기서, (1)에서는 문서 문장 처리 과정으로 정답 위치 색인 정보에서 특정 임계값 Θ보다 큰 n개의 문장들을 분석한다. (2)에서는 분석된 문장과 사용자 질의 사이의 관련 정도를 평가한다. 이렇게 평가된 각 문장의 관련 정도 중 가장 큰 값을 가지는 문장을 정답 문장으로 추출한다. 특히, (3)의 "w >= max_w"는 문서의 뒤에 있는 문장이 더 정답에 가깝다는 전제에 대한 조건이다. 다음은 질의와 문장을 분석한 결과에 대해 QAU와 SAU 사이의 유사도를 평가하는 알고리즘이다.

여기서, 식(4)에서 TAU, TAND 그리고 TOR는 하나의 a_au에 대한 타입으로 불리언 연산자인지 AU인지를 나타낸다. 즉, QAU는 AU와 불리언 연산자의 후순위 형태(post_fix)로 표현된다. 다음은 QAU의 AU와 SAU 사이의 관련 정도를 평가하는 알고리즘이다.

```
q_au= au;
max_w =0.0;
for each s_au in SAU;
      w = au_matching(q_au, s_au); ..... (5)
      if(w > max_w) max_w = w;
for end;
return max_au;
```

여기서, (5)은 두 AU 사이의 관련 정도를 평가하는 함수이다. 이 관련 정도는 AU의 핵심 관점(analysis_point)와 주변 관점(auxiliary_point)의 매칭을 통해 다음과 같이 평가된다.

$$au\_matching(au1,\ au2) = ALPHA * an\_matching(au1,au2) + BETA * aux\_matching(au1,au2)$$
$$ALPHA + BETA = 1.0$$

k_1 = au1.analysis_point.keyword
k_2 = au1.analysis_point.keyword
$CNet(k_1, k_2) = e^{-0.2 \times DIST(k_1, k_2)}$
$an\_matching(au1, au2) = 0.5 \times CNet(k_1, k_2)$

$INTER\_AUX = au1.auxiliary\_point \cap au2.auxiliary\_point$
$UNION\_AUX = au1.auxiliary\_point \cup au2.auxiliary\_point$
$aux\_matching(au1,au2) = \frac{n(INTER\_AUX)}{n(UNION\_AUX)}, union\_aux \neq \phi$

여기서, $union\_aux \neq \phi$ 이고 ALPHA와 BETA는 두 관점에 대한 중요도를 나타내며 도메인에 대한 튜닝을 통해 결정된다. CNet(k1, k2)는 개념망에서 두 개념 용어들 사이의 거리를 통한 관련 정도를 계산하는 기본 함수로 두 용어의 거리가 0이면 1의 값을 가지며, 거리가 멀어질수록 0에 가까워지는 성질을 가진다.

Ⅳ. 실험 및 평가

이 장에서는 정답 검색 시스템의 평가의 정답 위치 추출 엔진에 대해 평가한다. 여기에 이용된 질의의 개수는 총 100개이다. 이 질의는 개념어 100개와 속성어 80개를 대상으로 생성하였다. 평가에 이용된 정답 문서 개수는 총 8,000개이다. 이 평가 절차는 먼저 사용자 질의를 통해 정답 문서를 검색한 다음, 검색된 정답 문서들 중 상위 3개에 대해 전역 정답(GA: Global Answer)와 지역 정답(Local Answer)를 각각 추출하였다. 즉, GA는 검색된 정답 문서들 각각으로부터 색인된 상위 3개의 문장을 기준으로 5개 정답 추출하였으며, LA는 검색된 정답 문서들 각각으로부터 색인된 상위 10개의 문장을 기준으로 5개 정답 추출하였다.

표 1. 상위 K 번째 정답 포함율
Tab 1. Upper Kth Answer Including Rate

K	Global Answer	Local Answer
1	0.71	0.72
2	0.87	0.79
3	0.91	0.88
4	0.95	0.88
5	0.95	0.89

<표 1>은 상위 K = 1, 2, ..,5 번째 정답 포함율을 나타내고 있다. GA의 경우 K=5일 경우 0.95의 높은 정답 포함률을 나타내고 있다. 또한, LA의 경우 0.89의 포함률을 나타내고 있다. 이 평가에서 추출될 정답의 개수는 5개 내외에서 수렴하고 있음을 알 수 있다. 또한, 이 5개의 정답은 QA TREC에서 제시하는 정답의 개수와 일치한다.

표 2. 역순위 평균 정답률 평가
Tab 2. Evaluation of MMR(MMR:Mean Reciprocal Rank)

Rank	Global Answer	Local Answer
1	0.66	0.68
2	0.78	0.73
3	0.79	0.74
4	0.80	0.75
5	0.81	0.75

역순위 평균 정답률(MRR)은 정확히 일치하는 정답의 순위가 r일 경우, 1/r의 가중치를 부여하는 평가방법으로 K 이내 정답 포함률 보다 엄격하며 TREC에서 이용하는 평가 방법이다. 즉, 첫 번째 검색된 정답이 정확히 일치했을 경우 1, 두 번째 검색된 정답이 정확히 일치했을 경우 1/2와 같은 방식으로 점수를 부여한다. 본 시스템은 r=3일 때, GA와 LA가 각각 0.79과 0.74의 값을 가진다.

Ⅴ. 결론

본 연구에서는 지식검색을 위해 개념 속성을 이용하여 사용자 질의에 가장 적합한 정답 문장들을 검색할 수 있는 정답 검색 시스템을 설계하고 평가하였다. 이 시스템은 정답 위치 색인 엔진, 정답 문서 검색 엔진 그리고 정답 추출 엔진으로 구성되어 있다. 여기서, 사용자는 특정 질의에 대한 정답 문서를 검색할 수 있으며, 동시에 이 검색된 문서들로부터 자신이 요구한 정답 문장들을 검색할 수 있다. 또한, 특정한 문서를 선택함으로써 그 문서에 포함된 정답 문장들을 검색할 수 있다. 이 기능은 검색된 많은 문서를 전부 사용자가 읽어보지 않더라도 자신이 요구한 내용이 검색된 문서에 포함되어 있는지를 알 수 있게 한다. 특히, 정답 추출 엔진은 비교적 높은 정답 제시 기능을 가지고 있음을 K 이내 정답 포함률과 역순위 평균 정답률을 통해 평가하였다.

본 논문에서 제안한 시스템은 웹 인덱스방식의 문제점인 색인의 부정확할 경우 원하는 정보를 찾을 수 없는 점과 너무나 많은 검색결과들을 보여줌으로써 실제로 원하지 않는 결과도 함께 보여 주는 점을 완화하여 사용자가 원하는 정답을 보다 정확하게 제시할 수 있다. 또한 제안한 시스템은

질의응답 시스템 같이 개체명 인식등의 복잡한 자연어 분석을 수행하지 않더라도 개념과 속성으로 색인 및 검색함으로써 질의응답시스템 정도의 정답률을 보이는 우수한 시스템이다.

참고문헌

[1] Baeza-Yates, R. and Ribeiro-Neto, B., Modern Information Retrieval, Chapter 13, Addison Wesley Longman 1999.

[2] Buttler, D., Liu, L., and Pu, C., "A Fully Automated Object Extraction System for the World Wide Web", In International Conference on Distributed Computing Systems, 2001.

[3] Buyukkokten, O., Garcia-Molina, H., and Paepche, A., "Accordion Summarization for End-Game Browsing on PDAs and Cellular Phones", In Proc. of the Conf. on Human Factors in Computing systems, CHI'01, 2001.

[4] Gerald Salton, Automatic Text Processing, Addison-Wesley publishing company, Massachusetts, 1988.

[5] Gu, X., Chen, J., Ma, W.Y., and Chen, G., "Visual Based Content Understanding towards Web Adaptation", In Second International Conference on Adaptive Hypermedia and Adaptive Web-based Systems (AH2002), Spain, pp. 29-31, 2002.

[6] Hovy, Eduard, Ulf Hermjakob, Chin-Yew Lin, Deepak Ravichandran, "Using Knowledge to Facilitate Factoid Answer Pinpointing", COLING 2002.

[7] J. Perez-Carballo and T. Strzalkowski, "Natural language information retrieval: progress report," Information Processing and Management, Vol. 36, pp.155-178, 2000.

[8] Moldovan, Dan, Adrian Novischi, "Lexical Chain for Question Answering", COLING 2002.

[9] Newby, G., "Information Space Based on HTML Structure", In The Ninth Text REtrieval Conference (TREC 9), pp. 601-610, 2000.

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

[11] Sanda Harabagiu, Marius Pasca, and Steven Maiorano, Experiments with "Open-Domain Textual Question Answering", In Proceedings of COLING-2000, Saarbruken Germany, August 2000.

[12] Yang, Y. and Zhang, H., "HTML Page Analysis Based on Visual Cues", In 6th International Conference on Document Analysis and Recognition (ICDAR 2001), Seattle, Washington, USA, 2001.

[13] 김미진, 박미성, 장혁창, 이상조, 최재혁, "고빈도어를 이용한 복합명사 색인어 추출 방안", 제 10회 한글 및 한국어 정보 처리 학술 발표 논문집, pp.121-129, 1998.

[14] 김수민, 백대호, 김상범, 임해창, "시소러스 범주를 이용한 질의응답시스템", 제12회 한글 및 한국어 정보처리 학술대회, pp.179-183, 2000.

[15] 권승학, "인터넷 검색엔진에서 사례기반 추론을 이용한 인터페이스 에이전트 설계", 한국컴퓨터정보학회논문지, Vo. 5, No. 2, pp.50-59, 2000.

[16] 명순희, "협동에이전트를 이용한 정보검색, 한국컴퓨터정보학회논문지, Vo. 5, No. 2, pp.43-49, 2000.

[17] 박미화, 원형석, 이원일, 이근배, "구문분석에 기반한 자연어 질의로부터의 불리언 질의 생성", 제 10회 한글 및 한국어 정보처리 학술 발표 논문집, pp73-80, 1998.

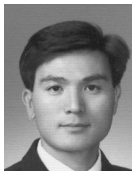
[18] 이경순,김재호, 최기선, "한국어 질의응답 시스템에서 개체인식에 기반한 대답추출", 제12회 한글 및 한국어 정보처리 학술대회, p.184~189, 2000.

[19] 이현아, "구문분석과 공기 정보를 이용한 개념 기반 명사구 색인 방법", 포항공대 전산과 석사 학위 논문, 1996.

[20] 장문수, 장명길, 김현진, 오효정, 이재성, "인터넷 질의/응답을 위한 지식베이스 구축" 제12회 한글 및 한국어 정보처리 학술대회, p.198~204

[21] 최대선, "구 색인에서 성분 단어의 가중치 부여 방법에 관한 연구", 포항공대 석사학위 논문, 1997.

저 자 소 개



윤 보 현

1999년 : 고려대학교 컴퓨터학과 이학박사

1999년~2002년 : 한국전자통신연구원 선임연구원(팀장)

2003년~현재 : 목원대학교 컴퓨터교육과 조교수

서 창 호

1996년 : 고려대학교 수학과 이학박사

1996년~2000년 : 한국전자통신연구원 선임연구원(팀장)

2000년~현재 : 공주대학교 바이오정보학과 부교수