제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003.5)

자연어 질의 처리 기반 지능형 정보검색

이은옥*, 이연식*
*군산대학교 전자계산교육학과
e-mail:lhy123@intizen.com

Natural Language Query Processing Based Intelligent Information Retrieval

Eun-Ok Lee*, Youn-Sik Lee*
*Dept of Computer Science, Kun-san University

요 약

웹 문서의 홍수 속에서 사용자의 요구에 맞는 문서만을 검색해 주는 정보 검색 시스템이 요구되고 있다. 자연어 질의를 이용한 정보검색 방법은 초보자도 사용이 쉽고 사용자의 의도를 파악하기가 쉬어 지능형 정보검색에 적합하다. 따라서 현재는 자연어 질의로부터 사용자의 의도를 파악하기 위한 다양한 연구가 진행되고 있다. 본 논문에서는 구조화된 자연어 질의에서 한국어의 문맥 구조를 기반으로 하여 사용자의 의도를 파악하고 이를 이용하여 정보검색 질의를 생성하는 방법을 제안한다. 이렇게 생성된 질의어를 이용해서 메타정보검색을 하면 보다 정확하고 사용자의 의도에 맞는 문서만이 검색되었다.

1. 서론

웹 문서가 기하 급수적으로 증가하면서 단순히 키워드를 이용하거나 불리언 질의를 이용해서 정보를 검색하게 되면 방대한 문서가 검색될 뿐만 아니라 이용가치가 없는 문서까지도 검색된다. 따라서 사용자가 원하는 정보만을 검색해 주는 지능형 정보검색에 대한 연구가 활발히 진행되고 있다. 지능형 정보 검색 방법 중의 하나인 엠파스[1]는 단순한 패턴매칭이나 명사구들의 추출에 의한 방법으로 아직도 불필요한 문서가 많이 검색되는 실정이다. 이를 개선하기 위해서 최근에는 자연어 질의를 제약하거나 사용자의 의도를 파악하기 위한 다양한 연구가 진행중이다.

본 논문에서는 한국어 질의의 문맥구조를 이용하여 질의어를 제약하는 방법을 기술한다. 예를 들면 "바다 물고기 중에서 생선인 조기에 대해 알고 싶다"라는 자연어 질의에서 명사들만을 추출하면 "바다 물고기, 생선, 조기"이다. 이렇게 추출된 키워드를 이용하면 "바다 물고기"나 "생선"에 관한 문서도 모두 검색된다. 하지만 사용자는 "조기"에 관련된 문서만을 요구한다.

한국어의 문맥 구조는 특성상 "NP1(체언구)에서 NP2(체언구)" 또는 "NP1인 NP2"의 형태가 많이 사용되며 NP1은 NP2의 상위 개념을 내포한다. 아울러 하위의 개념인 NP2는 상위의 개념도 내포하기 때문에 NP2만을 키워드로 추출해도 된다. 위의 한국어 질의는 "NP1에서 NP2인 NP3에 대해 알고 싶다"의 문맥구조를 가지므로 NP1과 NP2는 키워드로 추출하지 않고 NP3인 "조기"만을 키워드로 추출할 수가 있다. 이렇게 함으로써 많은 키워드를 줄일 수 있을 뿐만 아니라 정보 범람을 예방할 수가 있다.

또한 추출된 키워드 정보가 여러 가지 의미를 가지는 다의어를 포함하고 있으면 사용자가 원하지 않은 문서까지도 검색될 수가 있다. 위의 예에서 추출되는 키워드 "조기"는 다음과 같은 다양한 의미로 사용된다[2].

1) 참조기, 수조기 등을 일컫는 "굴비"를 의미

2) 반기(半旗)나 조의를 나타내기 위한 깃발을 의미
3) 조기 교육과 같이 이른 시기(早期)를 의미
4) 조기 축구와 같이 아침에 일어남(早起)을 의미
5) 조각하는 기술(彫技)
6) 기관이나 기계 등을 만드는 것(造機)을 의미
7) 낚시터(釣磯)를 의미

정보 검색에서는 1), 2), 3), 4)의 의미가 주로 사용되며 5), 6), 7)은 거의 사용되지 않는다. 그러나 사용자의 질의만을 이용해서 위의 의미를 구별하는 방법은 없다. 따라서 정보검색 시스템은 사용자의 정확한 의도를 알지 못한 채 위의 7가지 의미 모두를 검색하게 된다. 실제 예문에서는 1)의 의미를 가지는 문서만을 검색하면 된다. 이를 해결하기 위해서는 시스템과의 점진적인 대화를 통해서 사용자의 정확한 의도를 파악한 후에 그에 해당하는 동의어로 키워드를 대체하는 방법을 생각할 수가 있다.

따라서 본 논문에서는 한국어 질의가 가지는 문맥 구조의 특성과 시스템과의 점진적인 대화를 통해서 사용자의 요구를 보다 명확하게 표현할 수 있는 대화형 질의 처리 에이전트를 이용한 지능형 정보검색 방법을 제안한다.

2. 한국어 질의의 특성

한국어는 본질적으로 모호하지만 특정 분야에 한정되는 지식만을 이용한다면 모호성을 부분적으로 줄일 수가 있다. 예를 들면 정보 검색 시스템에서 사용되는 질의어의 문맥 구조를 파악하여 문장의 의미를 제약하면 보다 정확한 분석 결과를 얻을 수가 있다. 본 논문에서는 일반적인 자연어 질의를 모두 사용하지만 그 중에서도 다음과 같은 유형의 질의는 문맥 구조 질의 생성 시스템에 의해 문장의 의미를 제약한다.

가) NP1 중에서 **NP2에** 대해 알고 싶다
 - *생선 중에서 조기에 대해 알고 싶다*
나) NP1인 **NP2을/를** 찾아라
 - *생선인 조기를 찾아줘*
다) *NP1 중에서 NP2인* ***NP3***를 찾아라
 - *바다 생물 중에서 생선인 조기를 찾아라*
라) NP1에서 **NP2**을/를 찾아라/찾아줘
 - *바다 생물에서 조기를 찾아라*
마) 가) - 라)에서 용언이 생략된 유형
 - *바다 생물에서 조기*

위의 질의 유형에서 NP1과 NP2는 상위-하위 관계로 되어 있다. 대부분의 질의도 이를 바탕으로 이루어진다. 정보 검색을 하기 위한 키워드로는 NP2만을 이용해도 된다. 이런 유형을 따르지 않는 "사과, 배 중에서 가장 큰 과일은"처럼 NP1의 위치에 보다 하위 개념이 오게 되면 구문분석 결과만을 이용한다.

3. 대화형 질의 처리 에이전트

질의어에 "배"나 "조기"와 같은 다의어가 포함된 경우에는 많은 문서가 검색되게 된다. 이런 경우 질의어 정보만으로는 다의어의 의미를 정확하게 파악하기가 어렵다. 본 논문에서는 대화형 질의 처리 에이전트를 사용하여 시스템과 대화를 통해서 해결한다. 예를 들어 "조기"라는 키워드인 경우에는 앞에서 보인 1) ~ 4)의 의미를 보여주고 사용자로 하여금 선택하게 한다. 사용자가 선택하면 그 의미를 가지는 다른 키워드로 대체한다. 1)을 선택했다면 "굴비"라는 키워드로 바꾼 다음 검색을 하면 사용자가 원했던 문서만을 검색하게 될 것이다. 이를 위해서는 다의어 사전이 필요하며 다의어가 가지는 의미와 그에 해당하는 대역어가 수록되어 있어야 한다.

4. 지능형 정보 검색 시스템

4.1 질의어 분석

형태소 해석 결과를 이용해서 한국어 질의를 불리언 질의로 변환하는 방법은 주어진 질의어에 대해 적절한 검색어로 신속하게 변환할 수 있지만 결과의 정확성은 저하된다. 본 연구에서는 문형 정보와 구문 형태소를 이용한 구문 분석[3]을 통해 문법적 관계까지 고려하여 질의어를 분석한다.

예를 들어 다음의 자연어 질의를 형태소 분석이나 부분적 패턴 정보에 의해 불리언 질의를 생성할 경우 아래와 같은 불리언 질의가 생성된다.

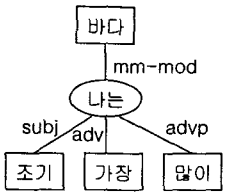
자연어 질의 : *조기가 가장 많이 나는* ***바다****는?*
형태소분석 후 불질의 : *조기 & 가장 & 나 & 바다*
올바른 불질의 : ***조기 & 바다***

이 질의에서 "가장"은 품사 모호성(명사, 부사)를 가지며 "나는"은 "나/대명사+는/조사"와 "날/동사+는/어미", "나/동사+는/어미"의 분석으로 인해 대명사인 "나"가 불리언 질의어의 키워드로 선택된다. 이러한 모호성은 질의 문장에서 "가장"과 "나는"의 문

법적 역할을 분석해야만 해결 가능하다. 따라서 사용자의 질의에 대한 정확한 분석을 위해서는 구문 분석 과정이 필요하다.

4.2 키워드 추출

키워드를 추출하는 방법은 다음과 같다. 질의어의 구문분석 결과인 파스트리를 순회하며 단말노드인 경우 범주(category)정보를 검사하여 보통 명사나 고유 명사 또는 이들을 포함하는 체언구로 분석된 경우 키워드로 추출한다. 관형격으로 사용된 비단말 노드인 경우는 후행하는 체언구를 키워드로 추출한다. 예로 "조기가 가장 많이 나는 바다는"의 경우 파스트리는 [그림 1]과 같고 키워드는 "조기, 바다"가 추출된다.



<그림 1>파스 트리

4.3 연산자 결정

입력된 질의어를 분석하여 키워드를 추출하면 키워드들 사이의 연산자를 결정해야 한다. 키워드들 사이의 연산자는 명사에 붙는 조사나 관형형 어미, 부사격 조사, 접속사의 종류에 의해 연산자를 결정할 수가 있다. 접속사가 존재하지 않는 경우는 기본적으로 'AND' 연산자를 부여하며 "정보나 검색"의 경우에는 'OR' 연산자를 "정보 그리고 검색"에서는 'AND'연산자를 부여할 수가 있다. 이와 같이 조사나 어미에 의해 연산자를 결정하는 것은 KT QUERY SET1.0[4]을 분석하여 정했으며 연산자의 유형은 <표 1>과 같다.

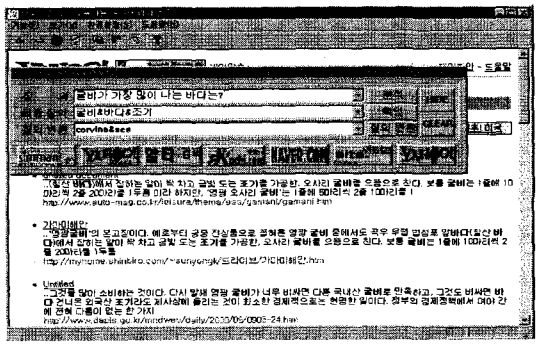
OR	AND	ANDNOT
	A와 B // A의 B	A를 제외한
	A B // A, B	A가 제외된
	A 그리고 B	A가 아닌
A 나 B	A 방식의 B	A를 포함하지 않는
A 거나 B	A를 이용한 B	A가 포함되지 않는
A 혹은 B	A를 위한 B	A를 뺀
A 또는 B	A에 사용되는 B	A가 빠진
	A (분야) 중 B	A 이외의
	A에 대한 B	A를 갖지 않는
	A 및 B // A 또 B	A가 들어있지 않는

<표 1> 불리언 연산자

5. 실험 및 평가

본 시스템은 윈도우 환경에서 작동 가능하도록 C 언어를 사용해서 구현한 본 시스템은 질의 처리 에이전트를 이용해서 사용자의 한국어 질의를 분석하여 사용자의 요구에 적합한 불리언 질의를 생성한 후에 현재 인터넷상에서 이용 가능한 다양한 검색엔진을 호출하여 검색 결과를 보여주는 기능을 한다. 실험 문장은 KT-SET[4]에서 200개와 자체적으로 수집한 영화 관련 질의어 200개를 사용하였다.

다음 [그림 2]는 질의어 "굴비가 가장 많이 나는 바다는?"을 입력하고 "분석"을 누른 결과이다. "질의변환"은 질의어의 확장이 필요할 때 사용한다. 예를 들어 "굴비"는 "조기"로 더 많이 사용되므로 "굴비 and 조기"로 질의확장을 할 수가 있다.



<그림 2>지능형 정보 검색의 실행

평가는 질의어에서 명사만을 추출하여 AND 연산자로 결합한 불리언 연산자를 임의로 만들어 네이버[5]와 비교하고 엠파스는 자연어 질의를 부분적으로 수용하기 때문에 주어진 질의를 원문대로 사용하여 비교하였다. 실험 결과, 본 논문에서 제안한 방법은 불리언 질의를 이용한 네이버보다 평균 18.72%의 성능이 향상되었고 엠파스를 이용한 방법보다는 20.32%의 성능 향상이 있었다. 이는 본 시스템에서는 질의의 확장 및 한정이 이루어졌으나 네이버나 엠파스는 질의의 확장이 없기 때문이다.

6. 결론

기존의 검색엔진은 불리언 질의나 부분적인 자연어 질의만을 이용했기 때문에 일반 사용자가 원하는 정보만을 쉽고 빠르게 검색하는 데는 한계가 있었다. 따라서 본 논문에서는 사용하기 쉽고 사용자

가 원하는 정보만을 검색해 주는 시스템을 제안했다. 본 논문에서 제안한 시스템은 불리언 질의가 아닌 자연어 질의를 기본 입력형태로 하였기 때문에 초보자라도 쉽게 사용할 수 있으며 정보검색 엔진마다 불리언 연산자가 조금씩 차이가 있었지만 자동으로 이를 생성해 주므로 사용자는 불리언 연산자를 알 필요가 없을 뿐만 아니라 검색엔진에 제약을 받지 않아도 된다. 또한 대화를 통해 사용자의 의도를 정확히 파악하기 때문에 사용자가 원하는 문서만 검색 될 뿐만 아니라 불필요한 문서의 검색을 방지하여 검색 효율을 향상시킬 수 있었다.

참고문헌

[1] 검색엔진 엠파스, http://www.empas.com

[2] 동아 새국어 사전 제 3판, PP.2004, 두산동아, 1999.

[3] H.Y. Lee, Y.G. Hwang, W.J. Bae and Y.S. Lee, "Unification Based Korean Parsing Using Sentence Patterns Information", NLPRS'99, pp.150-155, 1999.

[4] ktset95, http://nlp.korea.ac.kr/~cmj/kirs/cgi/ ktset.html

[5] 검색엔진 네이버, http://www.naver.com