論文
4-12-3

대화형 질의 처리 에이전트를 이용한 지능형 정보검색 (Intelligent Information Retrieval Using Interactive Query Processing Agent)

이현영(Hyeon-Yeong Lee)[1], 이기오(Gi-O Lee)[2], 한용기(Yong-Gi Han)[3]

요 약

대부분의 상업용 정보검색 시스템은 사용자의 질의 형태로 불리언 질의를 채용했다. 불리언 질의는 빠른 검색을 필요로 하는 검색엔진에는 유용할지라도 불리언 연산자로 사용자의 요구를 정확하게 표현하기는 어렵다. 따라서 사용자에게 편리한 자연어 질의를 이용하는 검색 엔진에 대한 연구가 있어왔다. 문서를 검색하기 위해서 사용자는 자신의 요구를 정확하게 표현해야 하며 사용자의 요구도 적절해야 한다.

따라서 본 논문에서는 자연어를 이용한 대화형 질의 처리 에이전트를 제안한다. 이 에이전트는 사용자와 점진적인 대화를 통해 사용자의 요구를 정확하게 표현한다. 사용자가 자연어 질의를 입력하면 에이전트는 질의를 분석하고 적절한 키워드를 추출하여 불리언 질의어를 생성한다. 추출된 키워드가 동의어이거나 다의어이면 사용자와 대화를 통해서 키워드를 한정하거나 확장한다. 이렇게 함으로써 사용자의 요구를 보다 구체적으로 표현하여 시스템의 성능을 향상시킨다. 따라서 본 시스템은 정보검색에서 정확률을 향상시킬 수 있다.

Abstract

Generally, most commercial retrieval engines adopt boolean query as user's query type. Although boolean query is useful to retrieval engines that need fast retrieval, it is not easy for user to express his demands with boolean operators. So, many researches have been studied for decades about information retrieval systems using natural language query that is convenient for user. To retrieve documents that are suitable for user's demands, they have to express their demands correctly.

So, this thesis proposes interactive query process agent using natural language. This agent expresses demands concrete through gradual interaction with user. When users input a natural language query, this agent analyzes the query and generates boolean query by selecting proper keyword and feedbacks the state of the keyword selected. If the keyword is a synonymy or a polysemy, the agent expands or limits the keyword through interaction with user. It makes user express demands more concrete and improve system performance. So, this agent can improve the precision of Information Retrieval.

1)정회원 : 서해대학교 컴퓨터정보기술계열 겸임교수

2)정회원 : 서해대학교 컴퓨터정보기술계열 부교수

3)정회원 : 서해대학교 컴퓨터정보기술계열 교수

논문접수 : 2003. 11. 7.
심사완료 : 2003. 11. 17.

1. 서론

인터넷의 활성화로 정보의 바다라 불리는 인터넷에서 사용자가 원하는 정보만을 효율적으로 찾아주는 정보검색 시스템의 중요성이 크게 대두되고 있다. 기존의 정보검색 시스템은 단순한 키워드나 키워드를 이용한 불리언 질의를 바탕으로 필요한 문서를 검색하기 때문에 방대한 양의 문서가 검색된다. 또한, 일반 사용자는 자신이 원하는 정보를 불리언(boolean) 형태의 질의로 표현하는데 익숙지 않으며 불리언 질의에 사용되는 연산자도 검색 시스템마다 차이가 있어 초보자가 사용하기에는 너무 어렵다. 따라서 최근에는 일상 생활에서 사용되는 자연어를 이용한 질의 방법이 연구되고 있다[1,2,8].

자연어로 질의하는 경우는 질의언어의 복잡한 구조에 상관없이 일상 대화처럼 질의할 수 있기 때문에 사용자의 편의성을 높일 수 있으며, 좀 더 정확한 질의로 변환할 수 있어 사용자의 질의 요구를 충실히 반영할 수 있다. 그러나 사용자들이 일반적으로 사용하는 2-3 어절의 짧은 질의는 사용하기에는 편리하지만 자신의 요구를 정확하게 표현하지 못한다. 그래서 사용자가 원하는 문서뿐만 아니라 원하지 않는 문서까지 검색하게 된다. 다음은 컴퓨터 과학과 정보과학 관련 분야의 질의어 모음[3]에서 발췌한 질의 형태이다. 사용자가 불리언 질의인 Q1이나 Q2의 형태로 질의하는 것과 자연어 질의인 NQ1이나 NQ2로 질의하는 것의 차이점을 명확히 볼 수 있다.

NQ1: 휴대용 컴퓨터나 노트북에 대해
Q1: 휴대용 컴퓨터 | 노트북 & 대해
NQ2: MS와 IBM에 대해 다루어진 문서 중 삼성을 제외한 문서
Q2: MS & IBM ~삼성

불리언 질의에 기반한 키워드 검색이나 단순한 패턴 매칭 방법에 의한 정보검색 방법은 사용자의 자연어 질의를 충분히 반영하여 검색하는 방법이 아니기 때문에 검색 결과의 질이 만족할 만한 수준은 아니다. 따라서 본 연구에서는 사용자가 원하는 정보를 검색하기 위해서 자연어로 된 질의를 입력하면 형태소 분석과 구문분석을 거쳐 적절한 키워드를 추출한 후에 이를 정보검색 시스템에 맞는 불리언 질의어로 변환해 주는 질의 처리 시스템을 제안한다. 또한 추출된 키워드에 여러 가지 의미를 가지는 동음이의어가 포함되어 있는 경우에는 사용자와 점진적인 대화를 통해 사용자의 의도를 파악해서 적절한 키워드로 대체하여 검색하는 지능형 정보 검색 방법을 도입한다. 또한, 사용자의 질의에 대한 최적의 불질의 생성을 위해 시소러스를 이용한 질의어 확장 및 제약을 통해 사용자의 요구를 충분히 질의에 반영할 수 있도록 한다.

결론적으로 본 연구에서는 시스템과의 점진적인 대화를 통해서 사용자의 요구를 보다 명확하게 표현할 수 있는 대화형 질의 처리 에이전트를 이용한 지능적인 정보검색 방법을 제안한다. 자연어로 된 질의어를 사용하며 자동으로 불리언 질의어로 변환해 주기 때문에 사용이 간편하고 이해하기가 용이했다. 또한 실험을 통해 키워드의 후보가 줄어들었고 결과적으로 검색되는 문서의 양이 줄어들었을 뿐만 아니라 검색된 문서의 정확성도 향상되었다.

2. 대화형 질의 처리 에이전트

2.1 연구배경

현재 사용중인 정보검색 시스템은 사용자의 질의를 "AND, OR, NOT"과 같은 논리 연산자를 사용하여 논리적 관계 표현인 불리언 질의로 변환하여 정보를 검색한다[4,5,6,7]. 그러나 대부분의 정보검색 시스템의 사용자들은 자신의 정보 요구를 불리언 질의로 표현하는데 익숙하지 않다. 또한 검색 엔진마다 불리언 질의로 변환하는 방법에 차이가 있다. 예를 들어 "종의 기원"에 대해 알고 싶을 때 사용자는 "종 기원"과 같이 키워드를 나열하여 정보를 검색하고자 한다. 그러나 사용자가 입력한 이 키워드는 심마니[4]에서는 "종 AND 기원"으로 변환되고 알타비스타[5]에서는 "종 OR 기원"으로 변환되어 검색된다. 이와 같이 불리언 연산

자의 의미나 검색 엔진의 특성을 모르는 일반 사용자가 자신이 원하는 정보를 정확하게 검색하기가 힘들다. 따라서 최근에는 사용자가 "종의 기원에 대해"와 같은 자연어 질의로 입력하면 이를 자동으로 불리언 질의로 변환해서 검색을 수행하는 연구[1,2,8]가 활발하다.

이러한 검색엔진은 자연어 질의를 자동으로 불리언 질의로 변환하기 때문에 사용자는 자신이 요구하는 정보를 정확하게 표현해야 한다. 그러나 사용자의 정보 요구를 정확하게 표현하기 위해서는 질의어가 길어지며 이 경우에는 이를 처리해야 하는 자연어 처리 시스템의 복잡도도 증가하게 된다. 또한 2-3어절의 짧은 질의를 사용한다면 검색 시스템은 사용자의 요구를 정확하게 파악하지 못하게 되어 원하지 않는 문서까지도 검색된다. 그러나 사용자들의 의도가 반영되지 못하는 2-3어절의 짧은 질의라고 해도 시스템과의 점진적인 대화를 통해서 사용자의 의도를 정확하게 파악한다면 사용자가 원하는 문서만을 쉽고 빠르게 검색할 수 있을 것이다. 즉 위에서 예를 든 "종 기원"에서 "종"이나 "기원"은 여러 가지 의미로 사용된다. 따라서 그 의미를 사용자에게 보여주고 그 중에서 자신의 요구에 맞는 의미를 선택하도록 제약하면 한가지 의미로 제약이 된다. 이렇게 함으로써 시스템은 그 의미를 가지는 키워드로 변환해서 검색을 한다면 사용자의 의도에 맞는 문서만을 검색할 수 있다.

따라서 본 논문에서는 시스템과의 대화를 통해서 사용자의 정보 요구를 정확하게 표현 가능한 "대화형 질의 처리 에이전트를 이용한 지능형 정보검색" 방법을 제안한다. 본 논문에서 제안하는 대화형 질의 처리 에이전트는 다음과 같은 특성을 가진다.

1) 불리언 질의 대신 자연어 질의를 이용하여 사용자의 편의성을 향상시킨다

2) 대화를 통해 사용자의 의도를 파악하기 때문에 원하는 정보만을 검색한다

3) 키워드의 확장과 한정이 자동으로 이루어져 불필요한 문서의 검색을 방지한다

4) 검색되는 문서의 양과 질의 향상으로 무선 인터넷 환경에도 적용할 수 있다.

2.2 불리언 질의 생성

2.2.1 질의어의 분석

사용자가 입력하는 한국어 질의 유형은 세 가지로 분류할 수가 있다. 대부분은 "종의 기원은"과 같이 용언이 생략되며 부가어의 수식을 받는 명사로 끝나는 형태[8]이며 "A는 무엇인가?", "A는 B한가?" 등의 의문문이거나 "A를 C하라" 등과 같은 명령문으로 구성된다[1]. 이들 질의어를 처리하는 방법에는 단순히 질의어를 형태소 분석하고 이를 바탕으로 명사만을 추출하여 키워드로 하는 방법과 일정한 패턴을 수집하고 수집된 패턴 내에서 불리언 질의 유형을 추출하여 부정확한 불리언 질의를 보완하는 방법이 있다. 이 방법은 구문 분석의 초기 단계인 부분 파싱 수준으로 이해될 수 있으며, 형태소 분석만을 이용하는 불리언 질의 생성에 비해 좀 더 나은 결과를 얻을 수 있다.

그러나 이들 방법은 한국어 단어가 가지는 품사 모호성과 하나의 단어가 다른 여러 형태소들의 결합으로 이루어질 때 나타나는 어휘 모호성으로 인해 불필요한 단어가 불질의의 결과로 나타나는 문제점이 있다. 예를 들어 "*굴비가 가장 많이 나는 바다는*"이란 사용자의 검색 질의에 대해 형태소 분석이나 부분적 패턴 정보에 의하여 불리언 질의를 생성할 경우, 아래와 같은 불리언 질의가 생성된다.

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

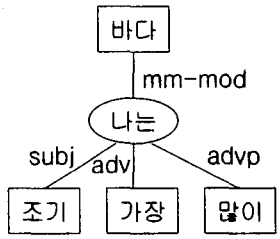
이 질의에서 "가장"은 품사 모호성(명사, 부사)을 가지며, "나는"은 "나/대명사+는/조사"의 분석과 "날/동사+는/관형형어미", "나/동사+는/관형형어미"의 분석으로 인해 대명사인 "나"가 불리언 질의어의 키워드로 선택된다. 이러한 모호성은 "가장"과 "나는"이 질의 문장에서의 문법적 역할을 분석해야만 해결이 가능하다. 따라서 사용자 질의에 대한 정확한 이해를 위해서는 부분 파싱 내지는 전체문장의 파싱이 반드시 필요하다.

2.2.2 키워드 추출

입력된 질의로부터 키워드를 추출하기 위해서는 한국어 처리 시스템을 사용한다. 형태소 분석과 구문분석을 거친 후 구문분석 결과인 구문 트리를 순회하면서 단말노드의 범주(category) 정보를 검사하여 유용한 키워드인지를 결정한다. 키워드가 될 수 있는 범주는 명사나 명사 상당어구(체언구)를 말한다. 전치사나 부사, 접속사 등은 키워드로 추출하지 않는다. 예를 들어 질의어 "조기가 가장 많이 나는 바다는"의 파스 트리는 <그림 1>과 같으며 이를 순회해서 얻어지는 키워드는 "조기"와 "바다"가 된다.

<그림 23>파스 트리

<Fig. 1>Parse Tree



2.2.3 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[3]을 분석하여 정했으며 연산자의 유형은 <표 1>과 같다.

<표 1> 불리언 연산자

<Table 1> Boolean operator

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

또한 연산자들은 질의어의 구문적 역할을 결정하는 질의 분석 결과의 비종단(Nonterminal) 노드에 의해 결정된다. 주격(SUBJ), 목적격(OBJ), 공동격(NP-WITH), 관형격(NP-MOD, MM-MOD)은 수식받는 체언구와의 관계에 따라 'AND' 연산자를 사용하며 조사가 "(이)나"이거나 접속사 "또는(NP-SET)"으로 연결된 경우에는 'OR' 연산자를 사용한다. 이러한 비종단 문법 심볼들은 문장에서 각 단어들 사이의 의존관계를 표현하고 있기 때문에 주어진 체언들을 제약하는 연산자를 쉽게 파악할 수 있다. 따라서 정확하고 올바른 불리언 질의 연산자를 생성하는 정보가 된다. 이런 유형의 연산자는 <표 2>의 구문정보를 이용한다.

<표 2>구문정보를 갖는 비종단 노드 유형
<Table 2>Nonterminal node with Syntactic Inforamtion

비종단 노드	의 미
SUBJ	주격
OBJ	목적격
NPADV	부사격
COOP	공동격
COMP	보격
ADVP	부사어에 의한 용언 수식
TIME	시간격
DEST	장소격
NP-MOD	'의'에 의한 체언 수식
MM-MOD	관형어에 의한 체언 수식
MA-MOD	부사어에 의한 체언 수식
NP-SEQ	조사가 없는 체언들의 나열
NP-SET	컴마(,)로 연결된 체언들
NP-WITH	'와'에 의한 체언들의 나열

2.2.3 대화를 통한 불리언 질의 생성

사용자가 입력한 자연어 질의를 분석하면 키워드를 추출할 수가 있다. 그러나 추출되어진 키워드가 여러 가지 의미를 가지는 다의어라면 사용자가 원하지 않는 문서까지도 검색될 수가 있다. 예를 들어서 "조기가 가장 많이 나는 바다는"이라는 질의어에서는 "조기"와 "바다"라는 키워드가 추출되었고 이를 불리언 질의로 변환하면 "조기 AND 바다"가 된다. 그러나 "조기"는 다음과 같이 다양한 의미로 사용된다 [9].

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 만드는것(造機)을 의미
7) 낚시터(釣磯)를 의미

정보검색에서는 1)과 2)가 하나의 체언구로 사용되고 3)과 4)는 뒤에 "축구"나 "교육"과 같은 체언이 뒤따르는 형태로 많이 사용되며 5), 6), 7)은 거의 사용되지 않는 의미이다. 즉 키워드 "조기"는 체언구가 바로 뒤따르면 3)이나 4)의 의미로 사용됨을 알 수 있다. 그러나 이러한 경우에도 사용자의 질의만을 이용해서 3)과 4)를 구별하는 방법은 없다. 따라서 시스템은 사용자의 정확한 의도를 알기 위해서 사용자와 대화를 시도한다. 즉, 위에서 나열한 의미들을 사용자에게 제시하면서 사용자가 의도한 의미를 선택하도록 하는 것이다. 이렇게 함으로써 사용자의 질의에 포함된 모호성을 제거하고 사용자의 정확한 의도를 파악하여 사용자가 원하는 정보만을 검색하기 위한 키워드를 추출한다.

2.2.4 불리언 질의 합성 모순 검사

사용자와의 대화를 통해서 불리언 질의를 생성하기 때문에 새로운 불리언 질의를 생성할 경우 전에 생성되어진 불리언 질의와 논리적인 모순이 존재하는지 검사해야 한다. 예를 들어 처음에는 "A에 대해 알고 싶어요"라는 질의어가 입력되었다면 불리언 질의는 "A"가 된다. 그런데 추가적인 대화를 통해서 "A는 제외시켜 주세요"라는 질의어가 입력된다면 새로운 불리언 질의어는 "A AND ~A" 또는 "A OR ~A"가 되어 논리적인 모순이 된다. 따라서 본 논문에서는 대화를 통해서 새로운 불리언 질의를 생성할 때 <표 3>을 참조하여 논리적인 모순을 해결한다.

<표 3>불리언 질의의 합성
<Table 3>Conjunction with Boolean Query

A and ~A	==> A
A or ~A	==> A
A and B and ~A	==> A and B
A and B or ~A	==> A and B
A and B and ~B	==> A and B
A and B or ~B	==> A and B
A or B and ~A	==> A or B
A or B or ~A	==> A or B

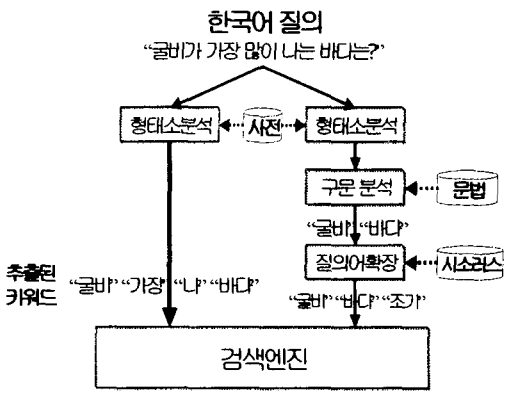
3. 한국어 대화형 질의 처리 시스템

인터넷상에서 정보를 검색하기 위한 한국어 대화형 질의 처리 시스템은 대화형 질의 처리

에이전트와 불리언 질의 생성 에이전트로 구성된다. 대화형 질의 처리 에이전트는 한국어 질의를 구문 분석하고 사용자와 대화를 하기 위한 시스템으로 한국어 전자 사전과 형태소 분석기, 구문 분석기로 구성된다. 불리언 질의 생성 에이전트는 구문분석된 결과를 이용하여 불리언 질의를 생성해 주는 시스템으로 질의어를 제약하거나 유사한 의미로 확장해 주는 질의어 확장과 질의 확장 정보로 사용되는 시소러스로 구성된다. 구문분석이 실패하는 경우에는 형태소 분석된 결과만을 이용한다.

<그림 24>대화형 지능형 에이전트를 이용한 시스템 구성도

<Fig. 2>Fig. of system using interactive query processing agent



<그림 2>는 한국어 질의에 대해 적절한 불리언 질의를 생성하는 과정을 설명한 것이다. 사전은 형태소 분석을 하기 위한 것으로 체언뿐만 아니라 용언에 대해서도 작성했으며 자주 사용되지 않는 의미나 고어는 생략했다. 이 사전은 "동아새국어사전 제3판"[9]에 나와 있는 약 16만 어휘로 구성되었다. 형태소 분석 과정에서는 사전에 등록되지 않은 고유명사도 추출할 수 있도록 미지어 추정 루틴을 포함하며 불용어 제거 루틴을 두어 키워드로서 가치가 없는 단어들은 제거한다. 구문 분석기는 한국어가 가지는 문법적 특징을 수용하면서도 빠른 구문 분석 결과를 가지도록 하기 위해서 문형 정보를 이용한 조건단일화 기반 GLR 파서[10]를 사용한다. 질의 확장 단계에서는 구문분석된 결과를 이용해서 정보검색에 필요한 키워드를 추출하며 추출된 키워드가 동음이의어이면 키워드 확장을 통해 다른 키워드로 대치한다. 또한 구문 정보를 이용하여 불리언 연산자를 생성하고 키워드와 연산자를 조합하여 불리언 질의를 생성한다. 이렇게 생성된 불리언 질의어는 정보 검색 엔진의 입력 정보로 사용된다.

4. 실험 및 평가

Windows나 UNIX 환경에서 동작하도록 C언어를 이용하여 작성된 본 시스템은 질의 처리 에이전트를 이용해서 사용자의 한국어 질의를 분석하여 사용자의 요구에 적합한 불리언 질의를 생성한 후에 현재 인터넷상에서 이용 가능한 다양한 검색 엔진을 호출하여 검색 결과를 보여주는 기능을 한다. 사용자가 한국어 질의를 입력하면 형태소 분석 결과와 구문 분석 결과를 선택적으로 볼 수 있으며 입력된 질의어가 의미적 중의성을 가지면 사용자와 대화를 통해 적절한 의미로 제약할 수 있도록 구성되었다.

4.1 질의 문장과 평가 방법

실험 및 성능 평가를 위해서 ETRI-KEMONG SET[11]의 46가지 질의어와 KTSET[3]에서 200개, 그리고 자체적으로 수집한 영화 관련 질의어 200개를 사용하였다. 그 중에서 일부를 보이면 다음과 같다.

<표 4> 질의 set

<Table 4> Query Set

상대성 이론의 허와 실은
당나귀와 말의 차이는
굴비가 가장 많이 나는 곳은
추석의 기원은
현재 상영되고 있는 영화는

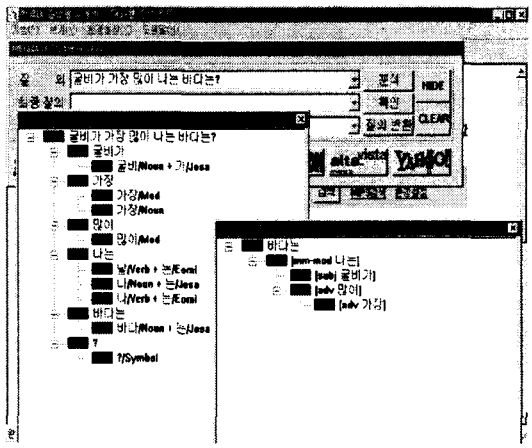
본 시스템은 초보자가 사용하고 있는 검색엔진을 보다 쉽고 효율적으로 사용하는데 목적

이 있다. 따라서 현재 상용 검색 엔진 중에서 자연어를 직접 질의로 입력할 수 있는 엠파스[7]와는 직접 자연어 질의를 이용해서 비교하고 자연어 질의를 입력할 수 없는 네이버[6]는 질의어에서 명사만을 키워드로 추출하여 불리언 질의를 임의로 생성하여 비교하였다. 비교 대상은 상위 20개의 문서만을 대상으로 사용자의 질의와 관련이 있는지 아닌지를 비교 분석하였다.

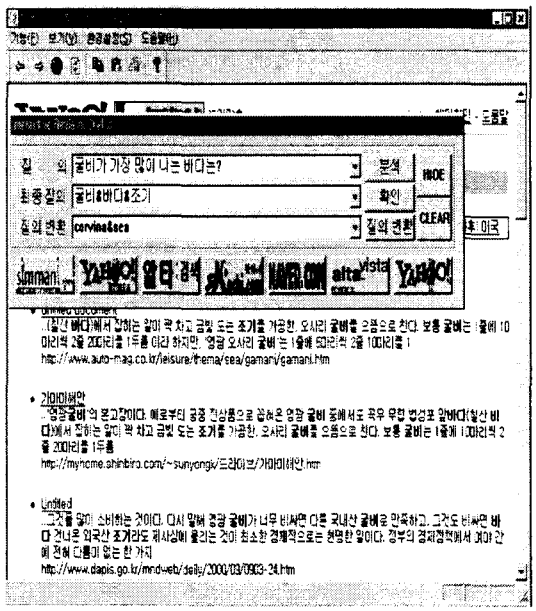
4.2 실험

<그림 3>은 사용자가 자연어 질의를 이용하여 정보검색을 하는 과정을 보여주는 예이다. 예를 들어 "굴비가 가장 많이 나는 바다는?"이라는 한국어 질의어를 입력하고 "분석"을 누른 경우로 입력된 질의어에 대한 형태소 분석 및 구문 분석 결과를 보여주고 있다. 여기에서 "HIDE"라고 되어있는 단추를 누르면 이 과정이 생략된다. "CLEAR"은 중간 결과를 제거하는 것이며 "질의 변환"은 질의어의 확장이 필요할 때 사용한다. 예를 들어 "굴비"는 "조기"로 더 많이 사용되므로 "굴비 and 조기"로 질의확장을 할 수가 있다. <그림 4>는 "질의 변환"을 눌러 불리언 질의를 확장한 경우이다. 생성된 불리언 질의를 이용해서 <그림 4>에 있는 검색 엔진을 클릭하면 해당 검색 엔진 사이트로 이동하며 검색 결과가 표시된다.

<그림 3>질의어의 형태소 분석 및 구문 분석
<Fig. 3>Morphological and Syntactic analysis of Query



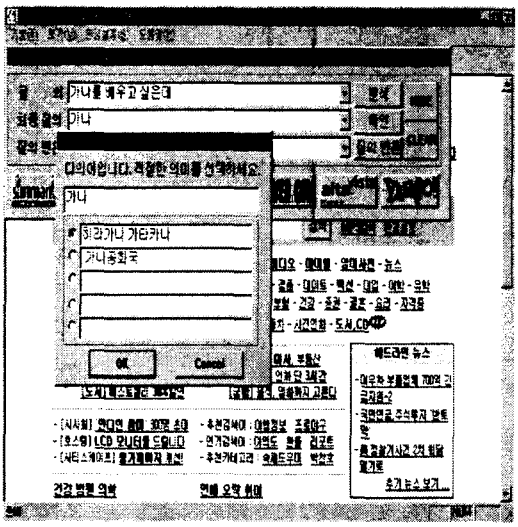
<그림 4>불리언 질의의 확장
<Fig 4>Expand of Boolean query



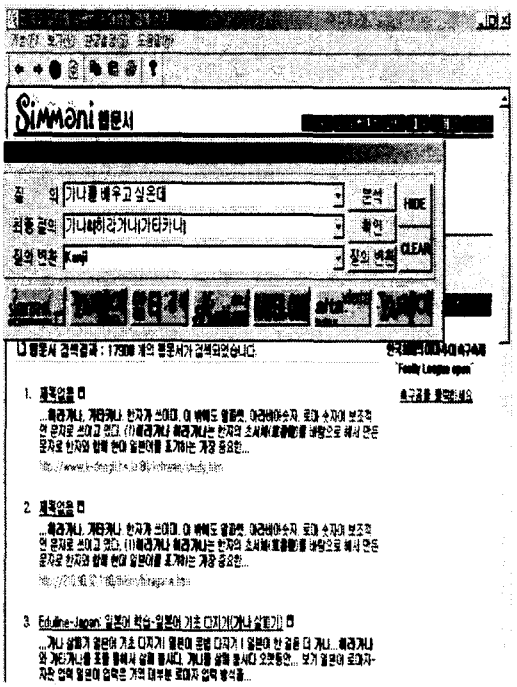
<그림 5>에서는 만약 질의어에 다의어가 포함된 경우 구축된 시소러스를 이용하여 자동으로 다의어에 제약이 가해지는 경우를 보여주고 있다. 질의 "가나를 배우고 싶은데"에서 "가

나"라는 검색어가 두 가지 의미를 가지고 있으므로 사용자와 대화를 통해 적절한 제약을 가하는 경우이고 <그림 6>은 제약이 가해진 결과를 통해 정보검색을 수행한 결과를 나타낸다.

<그림 5>시소러스를 이용한 질의어 제약
<Fig. 5>Limit of Query using Thesaurus



<그림 6>시소러스를 이용한 검색 결과
<Fig. 6>Retrieval Result using thesaurus



4.3 평가

평가는 두 가지 방법으로 행했다. 첫 번째 방법은 불리언 질의만을 사용하는 네이버와 본 시스템을 비교하였다. 네이버의 불리언 질의는 주어진 질의어에서 명사만을 추출하여 AND 연산자로 결합한 불리언 연산자를 임의로 만들었다. 두 번째 방법은 본 시스템과 엠파스의 검색 결과를 비교하였다. 엠파스의 경우도 자연어 질의를 부분적으로 수용하기 때문에 주어진 질의를 원문대로 사용하여 비교하였다.

실험 결과, 본 논문에서 제안한 대화형 질의처리 에이전트를 사용해서 정보를 검색하는 것이 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. 예를 들면 "굴비가 가장 많이 나는 바다는"에서 네이버는 "굴비 and 바다"로 질의를 하지만 본 시스템은 "굴비 and 바

다 and 조기"와 같이 질의 확장이 일어난다. 따라서 본 시스템은 네이버에 비해 "굴비"와 관련된 문서가 많이 검색되었다. 또한 엠파스에서는 "가장"처럼 문장에서 부사 역할을 하는 단어들도 키워드로 사용된다. 따라서 본 시스템보다 성능이 떨어질 뿐만 아니라 네이버보다는 성능이 떨어짐을 알 수가 있었다. 결론적으로 본 논문에서 제안한 사용자와의 점진적인 대화를 통해서 사용자의 정보요구를 보다 명확하게 표현하여 정보검색을 수행하면 정확률과 재현률이 높아짐을 알 수가 있었다.

5. 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 사용자가 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.

향후 연구과제로는 질의 확장을 위해서 사용하고 있는 유의어 사전과 국어 사전의 내용을 보충해야 하며 사용자의 편의성을 위해서 사용자가 입력한 질의어를 질의를 자동으로 확장할 수 있도록 키워드들간의 관련성 정보에 대한 연구가 필요하다.

참고 문헌

[1] 이승률, 강현규, 박세영, 이상조, "자연어 질의 정보 검색 시스템의 비주제어 탐색 방법을 통한 성능 개선", 제6회 한글 및 한국어 정보처리 학술대회, pp. 374-377, 1994.

[2] 이석호, 김성기, "자연 한글 질의어 처리를 위한 인터페이스의 설계 및 구현", 한국 정보과학회 논문지(C), Vol. 12, No. 1, pp. 31-44, 1985.

[3]KTSET95, http://nlp.korea.ac.kr/~cmj/kirs/cgi/ktset.html

[4] 검색엔진 심마니, http://www.simmani.com

[5] 검색엔진 알타비스타 http://www.altavista.co.kr

[6] 검색엔진 네이버, http://www.naver.com

[7] 검색엔진 엠파스, http://www.empas.com

[8] 이용석, 한국어 질의어를 수용하는 다국어 정보 검색 엔진 개발, 정보통신부, 산학연 공동기술개발사업 최종 보고서, 1999.

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

[10] Hyeon-Yeong Lee, Yi-Gyu Hwang, Woo-Jeong Bae and Yong-Seok Lee, "Unification Based Korean Parsing Using Sentence Patterns Information", NLPRS'99, pp.150-155, 1999.

[11] 한국전자통신연구원, ETRIKRMONG SET, 한국전자통신연구원, 1997.

한용기

1986년 전북대학교 전산통계학과 졸업(학사)
1988년 전북대학교 컴퓨터과학과 대학원 졸업(이학석사)
1996년 전북대학교 컴퓨터과학과 박사과정 수료
1991년-현재 서해대학 컴퓨터정보기술계열 교수
관심분야 : 자연어처리, 정보검색, 인공지능

이기오

1990년 전북대학교 전산통계학과 졸업(학사)
1993년 전북대학교 컴퓨터과학과 대학원 졸업(이학석사)
1995년 전북대학교 컴퓨터과학과 대학원 졸업(이학박사)
1995년-현재 서해대학 컴퓨터정보기술계열 부교수
관심분야 : 한국어정보처리, 음성인식, 자연어처리,정보검색

이현영

1991년 전북대학교 전산통계학과 졸업(학사)
1996년 전북대학교 컴퓨터과학과 대학원 졸업(이학석사)
1998년 전북대학교 컴퓨터과학과 박사과정 수료
2001-현재 서해대학 컴퓨터정보기술계열 겸임교수
관심분야 : 자연어처리, 정보검색, 에이전트시스템, 기계번역