자연어 질의 문맥 구조 기반 개인형 메타 검색 에이전트

박기선[0], 이덕남, 김우주, 이용석
전북대학교 {언어정보 공학실, 지능형 웹 & 전자상거래 연구실}
{kspark[0], dnlee}@cypher.chonbuk.ac.kr, {wjkim,yslee}@moak.chonbuk.ac.kr

A Personalizable Meta-Search Agent Based on Natural Query Context Structure

Ki-Seon Park[0], Deok-Nam Lee, Woo-Ju Kim, Yong-Seok Lee
Dept. of {Industrial Engineering, Computer Science}, Chonbuk National University

요 약

인터넷과 웹의 팽창과 함께 가용 정보의 양이 폭발적으로 증가하고 있으나 이에 대응되는 효과적이고 효율적인 정보 검색 능력의 지원이 없다면 이와 같은 방대한 정보들은 정보 이용자들에 있어 이용 가치가 없으며 이는 곧 정보 범람(information overflow)을 의미한다. 본 논문에서는 이에 대한 해결 방안으로써 사용자의 편이성과 정보검색 능력을 극대화할 수 있는 자연어 질의 문맥 구조 기반 개인형 메타 정보검색 엔진을 제안하고자 한다. 본 방법론은 자연어 질의를 기본 입력 형태로 하여 자연어 질의의 문맥 구조(context structure) 및 기타 정보 평가 요소들을 이용하는 다척도(multi-criteria) 의사 결정 기법 및 개인형 메타 정보 평가(information rating) 방법론으로 구성되어 있으며 이를 위한 시스템 설계를 제안한다.

1. 서 론

인터넷과 웹에서의 가용 정보가 폭발적으로 팽창함으로써 현존하는 국내외 검색 엔진들의 접근 방법으로는 정보를 검색하는 사용자가 원하는 정보를 쉽게 그리고 효과적으로 검색하기가 매우 어려워지고 있다. 즉시 이용할 수 있는 정보를 많이 갖고 있다고 해서 항상 양질의 정보는 아니다. 그것은 평가의 질을 낮추어 종종 의사 결정자들을 방해하곤 한다. 이는 많은 부분 현재의 검색 엔진들의 방법론의 한계점들로 인해 나타나는 현상들이다. 검색엔진으로부터 결과들이 불확실한 또 다른 이유는 사용자가 사용하는 용어의 의미와 검색엔진이 인식하는 의미적 차이인 "semantic gap" 때문이다. 그래서 이상적인 정보검색 시스템은 각 문서와 질의의 내용을 완전하게 이해하는 것이나 실제로 이것은 불가능하므로 대부분의 정보검색 시스템들은 문서들과 질의의 내용에 근접하는 어떤 구조화된 방법을 사용해야 한다.

Yahoo, Excite, Altavista, WebCrawler, Lycos, Google 과 같은 현재 대부분의 인터넷 검색 엔진들은 재현률과 정확률에 대한 문제점들을 외면하고 있다[1]. 몇몇 일반적인 검색엔진들은 질의의 정확성을 개선하는데 메타 검색 엔진들을 이용하려 한다. 예를 들면 MetaCrawler [2], SavvySearch [3], NECI 메타 검색 엔진 [4], Copernic (http://www.copernic.com) 등이 있다.

이 메타 검색 엔진들은 그럼에도 불구하고 재현률 문제에 미흡하게 대처하거나 부분적으로 정확률 문제를 다루는 방향으로 접근하고 있다.

본 논문에서는 자연어 질의 문맥 구조 기반 개인형 메타 검색 에이전트 접근 방법을 제안한다. 자연어 질의어에 나타나는 일정한 질의 유형을 파악하여 구조화된 질의 유형에 따른 주제어를 추출해 냄으로써 정보를 검색하는 사용자로 하여금 찾고자하는 정보가 상위에 랭크되거나 웹브라우저의 1 ~ 2 페이지에 랭크 될 수 있도록 하는데 목적을 둔다.

본 논문은 Scime와 Kerschberg[5,6]가 제안한 아이디어를 기반으로 하고 있다. 사용자들이 구체적으로 그들의 검색 성향을 구조화된 트리로 나타내는 기법을 제시하였다. 또한 다양한 구성요소들을 기반으로 하는 정교한 사용자 선호도 표현 스키마를 이용하여 의사 척도(decision-criterion)를 표현하였다. 웹 페이지의 검색 적중률을 평가하기 위해서는 NQCT(Natural Query Context Tree)와 구성요소 기반 선호도 표현법을 혼합한 의사 분석 기법을 사용하고 있다.

이미 잘 알려진 여러 정보검색 엔진들에서도 사용자가 필요한 정보를 검색했을 경우에 각 검색 엔진들의 특성에 맞게 정보를 검색해서 유용한 정보를 사용자에게 보여주는 데 상당한 성과를 보이고 있다. 하지만 사용자의 검색의지를 직접 구조화된 트리로 표현하는 것은 편이성 관점에서 매우 제한적이다. 따라서 본 논문에서는 자연어 처리 기법을 이용하여 구조화된 트리의 생성을 자동화함으로써 이러한 편이성 문제를 해결하고 동시에 향상된 검색 성과를 거두고자 한다.

본 논문에서는 이러한 자연어 질의 유형에 대해 알아보고 이를 메타 검색 시스템에 적용하여 여러 검색 엔진(Yahoo, Altavista, Naver 등)에서 수집해온 정보를 재순서화 하여 사용자가 원하는 정보가 상위에 랭크되는 개인형 메타 정보 검색 에이전트를 설계하고자 한다.

본 논문의 구성은 2장에서는 자연어 질의 유형에 대해 알아보고 3장에서는 자연어 질의 문맥 구조 트리 구조를 설명하고 4장에서는 본 논문의 시스템 구조를 제시하고 5장에서는 결론을 맺는다.

2. 자연어 질의 유형

개인형 메타 검색 에이전트에서 일반적인 자연어 질의를 이용할 경우에는 질의로서 의미가 없기 때문에 메타 검색 시스템에서 활용할 수 있는 구조화된 자연어 질의를 사용한다. 본 논문에서는 구조화된 자연어 질의 형태로 보편적인 것에 관한 질의 형태, 사용자가 해답을 요구하는 형태와 구체적인 자연어 질의 유형을 사용하였다. 개인형 메타 검색 시스템에서 구조화된 자연어 질의를 분석해서 개인형 메타 검색을 하기 위한 불리언 질의를 생성한다. 개인형 메타 검색 시스템에서 구조화된 검색 질의를 위해 사용되는 자연어 질의의 몇 가지 유형은 다음과 같다.

가) ~중에서 ~인(에 속하는) ~에 대해 알려줘.
 - 사무용 기기 중에서 사무용 가구인 의자, 책상에 대해 알려줘.
나) ~중에서 ~로는 ~를 ~으로는(로는) ~을 그리고 ~ 에 대해 알려줘.
 - 사무용 기기 중에서 사무용 가구로는 의자, 책상, 전화를 사무용품으로는 종이와 펜을 그리고 컴퓨터에 대해 알려줘.
다) ~중에서 ~를(을) 포함하는 ~를(을) 알려줘.
 - 사무용 기기 중에서 의자를 포함하는 사무용 가구를 알려줘.
라) ~중에서 ~내에 ~이 들어 있는 ~을 알려줘.
 - 사무용 가구 중에서 사무용품 내에 연필이 들어 있는 문서를 알고 싶어요.
마) ~를 제외한/포함하지 않는 ~을 알려줘.
 - 사무용품 중에서 파란색을 제외한 필통을 알려줘.
바) 가) - 라)에서 용언의 생략
 - 사무용품 중에서 종이와 펜을 포함하는 문서는.
사) 그 외
 - 사무용 가구에 대하여.
 - 책상과 전화에 관해.
 - 책상, 의자에 대하여
 - 사무용 가구와 사무용품의 차이는

3. 자연어 질의 문맥 구조 트리 기반 접근

3.1 자연어 질의 문맥 구조 트리

일반적인 키워드 기반 검색 표현은 사용자들의 검색 성향을 표현하기에는 불충분하다. 사용자들이 의사 결정(decision-making) 처리를 한다는 가정에 의해 쉽게 질의를 형식화 하여 검색을 지원 할 수 있다. 본 논문의 접근 방향은 한국어 문맥 구조의 계층적 개념 트리에 의해 사용자들이 질의어를 색인하고 사용자들의 검색 성향을 나타냈다. 검색에서 사용자들을 파악할 수 있는 검색성향의 개념을 반영한다. 이것을 자연어 질의 문맥 트리 모델(NQCT model)이라고 부른다.

예를 들면 "사무용 기기 중에서 사무용 가구인 의자와 책상에 대해 알려줘" 라는 문장에서 "~ 중에서 ~ 인 ~에 대해" 유형을 도출해 냄으로써 문장의 수식 관계로 인해 구조화[그림 1]를 시킬 수 있다.

```
[알려줘/알리(12) ASERV (*OR* QUES DEC)]
   +--LOC [사무용기기중에서/사무용기기(0)]
   +--NPADV [책상에대해/책상(8)]
         +--NP-WITH [의자와/의자(6)]
               +--MM-MOD [사무용가구인/사무용가구(1)]
```

[그림 1] 자연어 질의 문맥 구조를 분석한 결과 및 트리

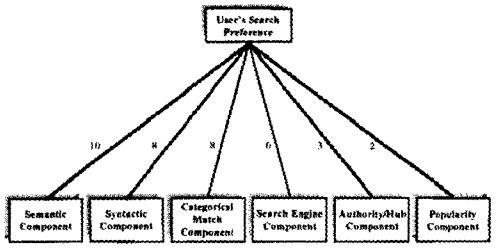
[그림 1]은 NQCT 스키마를 사용하는 사용자의 검색 성향의 현실적인 예이다. 사용자가 사무용기기와 사무용가구 문맥 내에서 의자, 책상에 대한 정보를 찾으려는 형태에 따라 [그림 1]의 서브 트리로 번역했으며, 이 번역된 트리에 기반을 둔 불리언 질의의 형태는 ((사무용가구&사무용기기&의자) | (사무용가구&사무용기기&책상))으로 조합되어 검색이 이루어진다.

3.2 다중 속성 기반 검색 선호도 표현

사용자에 대한 웹 검색 적중의 랭킹은 사용자 선호도와 결정문제의 개념을 반영하는 다중 속성의 평가를 포함한다. 본 논문의 접근에서는 다중 속성 결정 문제로서 랭킹의 문제를 제시한다. 이와 같이 다중 결정 척도(multiple decision criteria)에 따른 페이지 순위와 다중 검색엔진에 의해 제공되는 검색들의 결과를 실험할 것이며 다중 속성 유틸리티 기반[MAUT][7]과 Repertory Grid[8]로 정보 평가 문제를 설명할 것이다. 본 논문의 랭킹 접근은 MAUT와 Repertory Grid를 혼합한 형태이며 다음과 같은 6가지 검색 평가 구성요소들을 정의한다.

(1) Semantic component : 내용에 관련하여 웹 페이지 검색 능력을 나타낸다.
(2) Syntactic component : URL에 관련하여 구문 검색 능력을 나타낸다.
(3) Categorical Match component : 사용자가 생성한 분류 구조와 검색된 웹 페이지에 대한 검색 엔진들에 의해 제공된 카테고리 정보 사이에 측정된 유사도를 나타낸다.
(4) Search Engine component : 검색 엔진들 결과들에서 신뢰와 사용자들의 선입관을 나타낸다.
(5) Authority/Hub component : 권한 또는 허브 싸이트들과 페이지들[25]에 대한 사용자 선호도의 레벨을 나타낸다.
(6) Popularity component : 인기 있는 싸이트들에 대한 사용자들의 선호도를 나타낸다.

더욱이 이 다중 구성요소 기반 선호도 표현법 스키마에서 사용자는 검색엔진 구성요소 내에서 각각 사용 가능하고, 이 구성요소 각각에 선호도 레벨을 할당한다. [그림 2]는 개념적으로 위의 스키마를 묘사하고 있다. 모서리에 할당된 각각의 숫자는 그 구성요소에 대한 사용자들의 선호도 레벨을 보여준다. 이 다중 구성요소 선호도 스키마는 웹 페이지들의 검색 능력을 결정하고 검색을 통해 사용자들이 제어하는 것을 허용한다.



[그림 2] 사용자들 선호도 표현 스키마의 개념 모델

3.3 검색 성향을 기반으로 하는 웹 정보 수집

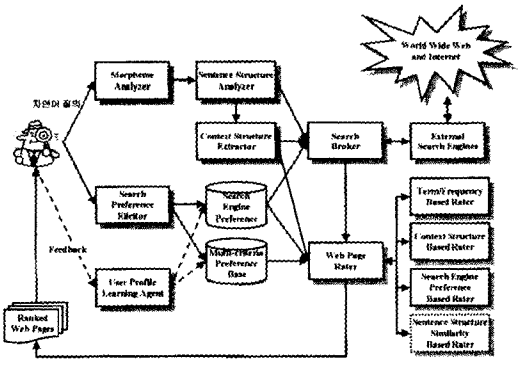
현재, NQCT에 기반을 둔 검색 요청을 받아들이는 검색엔진은 없다. 대부분 현재의 정보검색 엔진들이 처리할 수 있는 불리언 질의를 NQCT에 기반을 둔 질의로 번역하는 기법을 개발하여 각각의 현존한 검색엔진들이 받아들일 수 있는 분리된 질의 집합 내에서 [그림 1]과 같이 전체를 NQCT 트리로 변형한다. 이것을 위해서, 첫 번째로 루트에서 각각의 잎 노드까지 경로들의 집합 내에서 트리를 분리했다. 각각의 경로에 대해서 루트에서 잎 노드까지의 경로에서 각 노드의 긍정 개념 용어들로부터 하나의 용어가 선택되었을 때, 용어들의 모든 가능한 결합들을 생성한다. 마지막으로, 얻어진 질의 결과들을 검색엔진들에게 각각의 질의를 할당한다.

3.4 통합된 웹 정보 평가 기법

생성된 질의 문장에 대해 목표 엔진들로부터 각각의 결과 페이지 적중률을 각각의 구성요소에 대하여 평가를 한다. 각 웹 페이지의 6가지 검색 능력을 계산한다. 이 6가지 검색능력 값들의 합성은 다중 속성 기반 검색 선호도 표현 스키마의 기능에 기반을 두고 계산을 한다. 이 기법을 통해서, 각각의 웹 페이지는 사용자들의 관점으로부터 검색 능력의 레벨을 표현한 자신들의 값을 갖고 있다[9].

4. 시스템 구조

이 절에서는 자연어 질의 문맥 구조에 기반을 둔 개인형 메타 검색 에이전트 시스템의 구조를 나타낸다. [그림 3]은 자연어 질의 문맥 구조에 기반을 둔 개인형 메타 검색 에이전트 시스템의 전체 구조와 구성요소를 보여주고 있다. 중요한 정보의 흐름들을 묘사하고 있다. 자연어 질의 문맥 구조에 기반을 둔 개인형 메타 검색 에이전트 시스템은 8개의 부시스템과 4개의 중요한 정보 저장소로 구성되어 있다.



[그림 3] 자연어 질의를 이용한 문맥 구조 기반 개인형 정보 검색 및 평가 시스템의 구조

본 시스템에서 자연어 질의어가 입력되면 형태소 분석기(morpheme analyzer)와 구문 분석기(syntactic analyzer)를 이용하여 질의어의 문장 구조를 파악하게 된다. 파악된 구문 구조 정보는 추가적으로 문맥 구조 추출기(context structure extractor)를 통해 문맥 구조 분석을 실시하여

그 결과 문맥 구조(context structure)를 얻게 된다. 예를 들어 정보를 검색하는 사용자가 "사무용 기기 중에서 사무용 가구인 의자와 책상에 대해 알고 싶다."라고 자연어 질의를 했다면 앞에서의 세 단계를 거쳐 다음의 [그림 1]과 같은 문맥 구조를 추출할 수 있게 되는 것이다. 이 그림의 의미는 궁극적 검색 대상은 "의자"와 "책상"이며 이러한 의자와 책상을 정보를 검색하는 사용자가 사무용 기기와 사무용 가구의 문맥상에서 검색하고자 함을 의미하게 된다. 이러한 방식의 문맥 구조 정보는 단순한 키워드 추출 만으로의 검색보다 보다 정교한 검색과 평가를 가능하게 한다. 이상에서 추출된 정보들을 바탕으로 역시 [그림 3]과 같이 메타 검색기가 주요 검색 엔진들(Yahoo, empas, Naver 등)이 이해 할 수 있는 형태로 질의를 생성하고 그 질의 결과를 수집하게 된다. 이러한 자연어 질의 처리 외에서 본 방법론에서는 검색 엔진들에 대한 검색자의 선호 표현, 키워드 기반 평가, 문맥 구조 평가 등 정보 평가 척도들에 대한 중요성 평가 등을 검색 선호 추출기(search preference elicitor)를 통해 정보를 검색하는 사용자가 표현할 수 있으며, 이를 바탕으로 웹 페이지 평가기(web page rater)는 수집된 정보들을 종합적으로 평가하게 된다. 이러한 웹 페이지 평가기는 내부적으로 전통적 키워드 기반 정보 평가, 문맥 구조 기반 정보 평가, 검색 엔진 선호도 기반 정보 평가, 나아가 구문 유사도 기반 정보 평가를 먼저 실행하며 이들 각각은 [그림 3]과 같이 Term/Frequency Based Rater, Context Structure Based Rater, Search Engine Preference Based Rater, Sentence Structure Similarity Based Rater 등에 의해 수행되게 된다. 이들 각각의 평가 결과들을 다시 다척도 의사결정 이론에 근거하여 종합 평가를 수행하게 된다. 이러한 평가 결과를 바탕으로 정보를 검색하는 사용자에게 추출된 정보를 제시하게 되며, 제시된 정보에 대한 정보 검색자의 자발적 혹은 자동화된 피드백 메커니즘을 통해 정보 검색자의 다양한 선호 체계를 학습하게 되며 이러한 학습을 통해 정보 검색자의 검색 선호 프로파일이 지속적으로 개인 검색자의 특징을 반영하고 보다 효과적인 검색 성과를 제고 시키게 되는 것이다.

5. 시스템 구현을 위한 실험 및 평가

본 논문에서는 자연어 인터페이스를 갖추지 않은 WebShifter라고 불리는 WordNet을 이용한 검색 에이전트를 통하여 실험을 하였다. 이는 자연어 처리 단계를 거치지 않고 키워드들 사이의 상·하위 개념을 WordNet을 통하여 얻게 되어 사용자가 수동으로 질의를 구성하게 된다. 영어에 있어서는 이렇게 WordNet을 통하여 키워드들 사이의 상·하위 개념 정보를 얻을 수 있으나, 한국어의 경우 부분적으로 이용 가능하지만, 완벽하지 않으므로 문맥 구조를 통하여 질의를 구조화 한다. 이러한 WebShifter를 통한 실험 및 평가의 결과는 [그림4],[그림5]와 같다.

Search Engines	Hit Ratio	Average Rank of Relevant Pages
WebSifter	30%	9.17
Copernic	5%	20.00
Altavista	20%	13.50
Google	20%	14.00
Yahoo	10%	14.00
Excite	20%	15.00

[그림 4] "chair"의 경우에 있어서의 검색 성능 비교

[그림4]에서 볼 때 WebShifter 접근법이 정확률과 연관문서에 대한 평균 랭크라는 두 가지 척도에 있어서 다른 접근법에 비해 보다 성능이 뛰어나다.

Copernic은 검색 엔진들의 평균에도 못미치는 성능을 보였다.

대부분의 검색엔진들에 있어서 연관 문서가 하위에 랭크되는 반면 메타 검색엔진은 상위에 랭크된 문서를 먼저 고려하게 된다.

Search Engines	Hit Ratio
WebSifter	95%
WebSifter (w/o Categorical Match)	80%
Corpenic	75%
Altavista	65%
Google	60%
Yahoo	85%
Excite	65%

[그림 5] "office" 와 "chair"의 경우에 있어서의 검색 성능 비교

[그림5]에서 볼 때 WebShifter가 categorical match를 고려하지 않고 여전히 Copernic 에 비해 성능이 뛰어났으며 categorical match를 고려할 경우 성능이 15% 향상되었다. 이러한 실험 및 평가 결과를 기반으로 하여 한국어 자연어 질의 문맥 구조 기반 개인형 메타 검색 에이전트를 설계한다.

6. 결 론

본 논문에서는 두 가지 중요하고 보완적인 목표들을 이루기 위해 자연어 질의 문맥 기반 개인형 메타 검색 에이전트 접근 방법을 제안한다. 첫 번째, 웹 검색을 형식화하여 사용자들이 더욱 강한 표현을 할 수 있도록 허용할 것이다. 두 번째, 사용자들의 실시간 성향에 기반을 둔 검색 결과들의 검색 능력을 개선할 것이다. 또한 자연어 질의 유형이 다양해야 할 것이며 문맥 구조를 파악하여 생성한 트리 형태에서 선택된 용어들에 대한 유의어나 다의어 해결 방법도 앞으로 모색되어야 할 것이다. 우선, 사용자들의 검색 성향과 강력한 선호도 표현을 향상시키기 위해서 자연어 질의의 문맥을 파악한 개념들의 구조화된 트리로서 결정 문제를 표현하고, 구체적인 도메인 명세화 개념에 의해 실제 검색 성향을 사용자들이 표현함으로서 자연어 질의 문맥 기반 트리, 검색 성향 표현 스키마를 제안한다. 또한, 6가지 선호도 구성요소들의 기능으로서 검색 결과 평가 선호도를 사용자가 표현할 수 있도록 허용한다. 두 번째로, 검색된 정보의 선호도를 향상시키기 위해서 NQCT에 의해 나타난 사용자들의 검색 성향과 다중 선호도 구성요소들에 의해 나타난 사용자들의 검색 선호도 둘 다를 고려한 하이브리드(hybrid) 평가 기법을 제안한다. 세 번째로, 웹 페이지 검색을 위해, 잘 알려진 엔진들과 자연어 질의에서 문맥 구조를 파악하여 함께 협력하는 자연어 질의 문맥 기반 개인형 메타 검색 에이전트 시스템 프로토타입을 설계하였다.

본 논문에서는 개별적인 페이지의 새로운 구성요소를 평가하는 사용자의 욕구에 따라 현존하는 평가 방법과 함께 고려될 것이다. 이것은 사용자의 변화 욕구와 필요성으로 구성요소의 선택에 의한 개인화를 증가시킬 것이다. 실험을 유효하게 하기 위해 시스템을 실세계 실험을 할 것이다.

[참고문헌]

[1] Lawrence, S. and Giles, C. L., "Accessibility of Information on the Web," Nature, vol. 400, 1999, pp. 107-109.

[2] Selberg, E. and Etzioni, O., The MetaCrawler Architecture for resource Resource Aggregation on the Web," IEEE Expert, vol. 12, no. 1, 1997,pp.11-14.

[3] Howe,A.E.and Dreilinger, D., "Savvy Search:A Metasearch Engine that Learns which Search Engines to Query," AI Magazine, vol. 18, no. 2, 1997,pp. 19-25.

[4] Lawrence, S. and Giles, C. L., "Context and Page Analysis for Improved Web Search," IEEE International Computing, vol. 2, no. 4, 1998,pp.38-46.

[5] Scime, A. and L. Kerschberg, "WebSifter: An Ontology-Based Personalizable Search Agent for the Web," International Conference on Digital Libraries: Research and Practice, Kyoto Japan, 2000, pp. 493-446.

[6] Scime, A. and L. Kerschberg, "WebSifter: An Ontological Web-Mining Agent for E-Business," Proceedings of the 9th IFIP 2.6 Working Conference on Database Semantics (DS-9): Semantic Issues in E-Commerce Systems, Hong Kong, 2001.

[7] Klein, D. A., Decision-Analytic Intelligent Systems: Automated Explanation and Knowledge Acquisition, Lawrence Erlbaum Associates, 1994.

[8] Boose, J. H. and J. M. Bradshaw, "Expertise Transfer and Complex Problems: Using AQUINAS as a Knowledge-acquisition Workbench for Knowledge-Based Systems," Int. J. Man-Machine Studies, vol. 26, 1987, pp. 3-28.

[9] Kerschberg, L., et al., "A Semantic Taxonomy-Based Personalizable Meta-Search Agent," Forthcoming in Lecture Notes in Artificial Intelligence, 2002.