메타 검색을 위한 한국어 질의 생성에 관한 연구

이덕남*, 이용석*
*전북대학교 컴퓨터공학과
e-mail : {dnleejin, yslee}@chonbuk.ac.kr

A Study on Korean Query Generation for Meta Retrieval

Deok-Nam Lee*, Yong-Seok Lee*
*Dept of Computer Science, Chonbuk National University

요 약

인터넷의 급속한 팽창으로 인해 가용 정보의 양이 폭발적으로 증가하고 있으나 이에 대응되는 효과적이고 효율적인 정보 검색 능력의 지원이 없다면 방대한 가용 정보들은 정보 사용자들에게 있어 이용 될 가치가 없으며 이는 곧 정보 범람(information overflow)으로 이어진다. 본 논문에서는 이에 대한 해결 방안으로써 한국어 표준 문형의 패턴을 기술하고 한국어 문장 구조(Korean Syntax Structure) 파악을 통한 메타 검색 시스템 설계를 제안한다.

1. 서론

대부분의 검색 엔진들은 키워드와 이들의 수평적 조합을 기반으로 정보를 검색하고 있다. 이러한 키워드 기반의 검색으로 정말로 정보 검색 사용자가 무엇을 검색하고자 하는 지에 대한 검색 의지를 정확히 파악하고 반영하기가 매우 어렵다. 이러한 어려움은 단일 키워드에 대한 복수 의미 내재, 키워드에 대한 문맥, 키워드 간의 정확한 관계 등이 무시되고 이를 정보 검색에 이용하고 있지 못하기 때문이다.

인터넷 정보 검색에서 사용자들이 주로 사용하는 질의는 2 - 3개의 용어로 이루어진 짧은 질의가 대부분을 차지한다고 보아도 과언이 아니다. 또한 동음이의어를 갖는 용어를 사용하기도 한다. 짧은 질의를 처리하는 일반적인 방법은 시소러스[8]나 WordNet[1]을 이용한 질의 확장이다.

그러나 시소러스, WordNet과 같은 지식 베이스는 구축하기가 용이하지 않으며, 도메인 종속적인 면과 단어의 희귀(sparseness) 문제를 극복하기 어려운 단점이 있다. 또한 동음이의어 용어로 인하여 검색의 정확성이 떨어지는 문제점이 있다.

짧은 질의의 질의 확장은 일반적으로 시소러스[8]나 WordNet[1]과 같은 지식 베이스를 이용하는 것이다. 이들 지식베이스에는 상위어(BT), 하위어(NT), 관련어(RT), 동의어들이 포함되어 있기 때문에 질의 확장에 유용하게 사용될 수 있다.

질의 확장을 위한 방법으로 사용자에게 질의 형식화 도구[7]를 제공하여 사용자 스스로 검색하고자 하는 내용을 형식화하도록 하는 노력이 있다.

이는 사용자의 정보 표현의 욕구를 반영하는 긍정적인 면도 있지만, 이러한 질의 형식화 도구 역시 시소러스를 근간으로 하였기 때문에 시소러스의 단점을 그대로 지니게 된다.

상호 정보를 정보 검색에 이용하려는 노력에는 [2][8][9]이 있다. [2]는 순위조정에 상호 정보를 이용하였고, [8]은 시소러스의 자료 희귀를 보완하기 위해 상호 정보를 활용하였다. [9]은 색인어의 의미를 고려하여 색인어 선정 작업에 상호 정보를 이용하였다. [2][8][9]은 상호 정보가 검색 시스템의 정확률을 향상시킬 수 있음을 보여준다.

본 논문에서는 인터넷에서 웹 문서를 효율적으로 검색하기 위한 한국어 문장 구조를 기반으로 하는 한국어 질의 생성 방법을 제안한다.

2. 한국어 문장 구조 표준 문형

한국어 문장은 그 쓰임에 있어 생략, 도치, 장문의 복합 명사 사용이 빈번하다. 특히 뉴스그룹이나 경매 시스템, 기존의 지식 정보 거래 시스템의 경우 통신 특유의 문장 형식들로 인해 의미 기반 정보검색 기법을 적용하기엔 큰 문제가 있다. 다음 예문의 경우 국내 특정 지식 정보 사이트의 질의문 중 한 예이다. 이러한 문장의 경우 언어처리를 통한 자동적 정보 중개는 불가능하다고 본다. 따라서 사용자의 언어 기술을 제약해야만 할 방법이 제시되어야 한다.

(1) 국내에서 westlife가 공연한 적이 있나요?
(2) 그룹 신화 오빠들 앨범 언제 나오나여?

문형 표준안의 구성[10]은 표준문형, 표준문형의 유도와 어휘부로 구성된다. 표준문형은 다시 기본문의 구성과 확대로 이루어졌으며 이에 대한 해석과 제한은 표준문형 유도에서 기술한다. 또한 어휘부에서는 표준화된 어휘 범주 분류에 대해 기술한다. 문형 표준안은 서술어에 따른 논항들의 위치를 고정하여 자유어순 문제에 접근하였고, 표준문형 유도에서 추정패턴을 이용하여 생략현상에 대한 접근방법을 제시한다.

2.1 기본문의 개념

기본문은 기본적인 문장패턴에 대한 규정으로서 필수 논항과 서술어로 구성되어 있다. 기본문의 구성체인 논항과 서술어는 일정한 순서를 지니고 있으며, 예외적인 경우 단서조항을 둔다. 문장은 최소한 하나의 주어와 서술어를 지녀야 한다. 다만, 문장이 문장 안에 내포된 경우나 연결어미를 이용하여 문이 접속될 경우 논항의 일부가 생략될 수 있다. 모든 서술어는 일정수의 논항을 요구하며, 동사, 형용사, 혹은 '명사, 대명사, 수사+이다'로 구성된다. 서술어는 부사어, 부사구, 부사절에 의해 수식을 받는다.

2.2 기본문의 분류

기본문은 서술어가 요구하는 논항의 수에 따라 1항 술어, 2항 술어, 3항 술어로 나뉜다. 각 술어에 의해 요구된 논항은 그 위치가 각각 정해져 있으며, NP1은 주어 자리, NP2는 목적어 혹은 보어 자리, NP3는 필수 부사어 자리이다.

논항1 : NP1(주어) -NP+주격조사
논항2 : NP2(목적어/술어) -NP+목적격조사/보격조사
논항3 : NP3(필수부사어) -NP+부사격조사

2.2.1 1항 술어

(문형1) S → NP1 VP -1항 술어

컴퓨터가 확인된다.	(1)1
컴퓨터는 편리하다.	(2)
컴퓨터는 기계이다.	(3)
*확인한다.	(4)
*확인하다. 컴퓨터가	(5)

예문 (1)은 동사를 서술어로 하며, 예문 (2)는 형용사, 예문(3)은 명사에 지정사 '이'가 결합되어 서술어 역할을 하는 예이다. 그리고 예문 (4)는 필수격이 생략되어 비문이며, 예문 (5)는 고정 순서 위반으로 비문이다.

2.2.2 2항 술어

(문형2) S → NP1 NP2 VP -2 항술어
(문형3) S → NP1 NP3 VP -2 항술어

xml은 문서구조를 나타낸다.(NP2)	(1)
xml이 희망이 되었다.(NP2)	(2)
*유니코드를 xml은 채택한다.	(3)
xml은 html과 비슷하다.(NP3)	(4)
html과 xml은 비슷하다.	(5)

예문 (3)은 고정 순서 위반으로 비문이다. 그리고 예문 (4)가 2항 술어인 반면, 예문 (5)는 'html과 xml은' NP1이 되는 1항 술어로 본다.

2.2.3 3항 술어

(문형4) S → NP1 NP2 NP3 VP -3 항술어
(문형5) S → NP1 NP3 NP2 VP -3 항술어

자연어처리는 언어학을 모태로 여긴다.	(1)
자연어처리는 검색에 편리성을 준다.(수여동사)	(2)
자연어처리는 언어학을 전산학적으로 말한다.(발화동사)	(3)
*자연어처리는 언어학으로 모태를 여긴다.	(4)

전술한 바와 같이 VP가 수여동사거나 발화동사일 경우 문형 4,5 모두 적용될 수 있다. 그러나 예문 (4)와 같이 수여동사, 발화동사 이외의 동사가 문형 (5)로 쓰였을 경우 비문으로 간주한다.

2.3 논항 확대

NP1, NP2, NP3는 앞에 관형어를 선행시키거나 다른 NP와 결합하여 논항을 확대할 수 있다. 관형어는 관형사, NP+의(관형격 조사) 그리고 관형절을 포함한다. 관형절에 의한 논항의 확대는 절에 의한 확대이므로 문의 내포에서 다룬다.

또 다른 확대의 방법은 접속 조사를 이용한 병렬구성, 접속 부사를 이용한 나열식 구성 그리고 병렬구성으로 확대된 논항이 접속 부사로 결합된 복합구성이 있다. 결구 각각의 논항 1, 논항 2, 논항 3은

아래와 같은 방법으로 확대될 수 있다.

새 버전 (1)
새 버전의 개발자 (2)
마이크로소프트와 IBM이 공동으로 개발하였다. (3)
basic, C, COBOL 그리고 Pascal은 프로그램 언어다. (4)
*나는 IBM 한국지사 서비스 사업본부에 문의하였다. (5)

위의 예문에서 (1)은 관형사에 의한 논항확대이고, (2)은 관형격 조사, (3)와 (4)는 각각 접속 조사와 접속 부사를 이용한 논항 확대의 예제이다. (5)의 경우는 다음의 NP 구성 규칙에 의해 비문으로 처리된다.

(규칙 1) NP → (nc|nn)+≤3|nb|np
(규칙 2) NP′ → NP

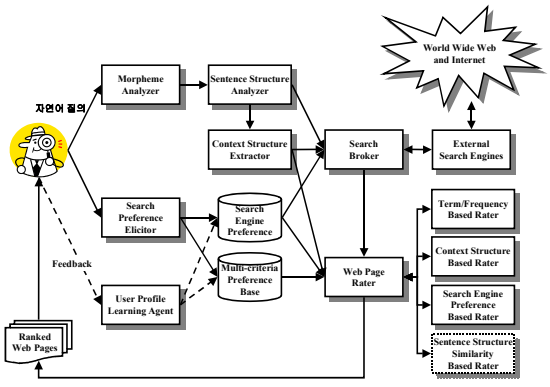
규칙 1은 NP 구성 규칙으로써 3 개 이하로 제약된 복합명사를 이루는 연속된 nc(일반명사)나 nn(수사), 또는 nb(의존명사), np(대명사)의 생성을 표현한다. 이후 NP는 규칙 1의 정의를 따른다. 아래의 예문에서 (1)-(4)는 정문이 되며 (5)는 비문으로 간주한다.

넷스케이프 버전 (1)
인터넷 익스플로러 설치 (2)
칠천 팔백 오십라인 (3)
애플릿 하나 (4)
*인터넷 익스플로러 설치 방법 (5)

규칙 2는 NP를 논항 확대된 NP′로 간주함을 나타내며, 이후 NP′는 논항확대를 표시한다.

3. 메타 검색 시스템 구조

(그림 1)은 본 연구에서 구축하고자 하는 "메타 검색을 위한 한국어 질의 생성 시스템"의 구조도를 나타낸다.

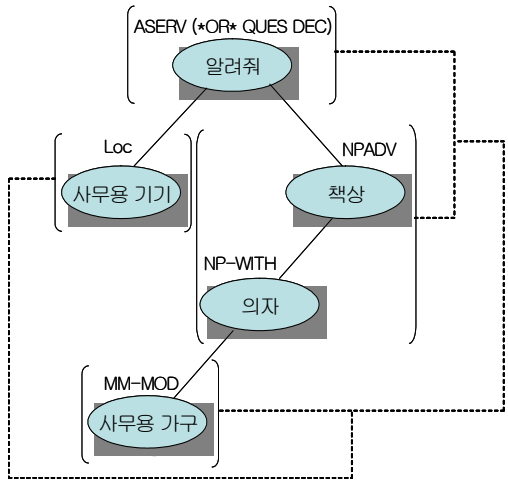


(그림 1) 한국어 메타 검색 질의 생성 시스템 구조

메타 검색을 위한 한국어 질의 생성 시스템의 진행 동작을 살펴보면 다음과 같다.

① 한국어 질의어 작성
② 한국어의 문장 구조 분석
③ 한국어 문장 구조 정보 추출(KSSE)
④ 메타 검색기를 통한 질의 결과 수집
⑤ 검색 이용자의 검색 선호도 추출
⑥ 웹페이지 평가(웹페이지 평가기)
⑦ 이용자의 검색 선호도 반영 프로파일 생성

예를 들어 정보 검색 이용자가 "사무용 기기 중에서 사무용 가구인 의자와 책상에 대해 알려줘." 라고 질의를 했다면 ①~③ 단계를 거쳐 (그림 2)와 같은 문장 구조 정보를 추출할 수 있다. 궁극적 검색 대상은 "의자"와 "책상"이며 이러한 의자와 책상 정보를 검색하는 사용자가 "~중에서 ~인 ~에 대해 ~"라는 한국어 문장의 구조 패턴을 고려한 사무용 기기와 사무용 가구의 문맥상에서 검색하고자 함을 의미하게 된다. 이러한 방식의 문장 구조 정보는 단순한 키워드 추출 만으로의 검색보다 보다 정교한 검색이 가능하다.



(그림 2) 한국어 문장 분석 세부 형태

한국어 질의 처리 외에 본 방법론에서는 검색 엔진들에 대한 정보를 검색하는 사용자의 선호 표현, 키워드 기반 평가, 문맥 구조 평가 등 정보 평가 척도들에 대한 중요성 평가 등을 검색 선호도 추출기(search preference elicitor)를 통해 검색 사용자의 검색 선호도를 표현할 수 있으며, 이를 바탕으로 웹페이지 평가기(web page rater)는 수집된 정보들을 종합적으로 평가하게 된다.

웹 페이지 평가기는 내부적으로 전통적 키워드 기반 정보 평가, 문맥 구조 기반 정보 평가, 검색 엔

진 선호도 기반 정보 평가, 나아가 구문 유사도 기반 정보 평가를 먼저 실행하며 이들 각각은 (그림 1)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 의사 결정 이론에 근거하여 종합 평가를 수행하게 된다.

앞서 평가된 결과를 바탕으로 정보를 검색하는 사용자에게 추출된 정보를 제시하게 되며, 제시된 정보에 대한 정보를 검색하는 사용자의 자발적 혹은 자동화된 피드백 메커니즘을 통해 정보를 검색하는 사용자의 다양한 선호 체계를 학습하게 되며 이러한 학습을 통해 정보를 검색하는 사용자의 검색 선호 프로파일이 지속적으로 정보를 검색하는 개인 사용자의 특징을 반영하고 보다 효과적인 검색 성과를 제고 시키게 되는 것이다.

4. 결 론

이미 Yahoo, Google, AOL 등 많은 외국 검색 엔진들은 검색 정보들에 대한 카테고리 즉 문맥 정보를 제공하고 있으며 국내의 한글 야후, 엠파스, 네이버 등의 검색 엔진들도 유사 문맥 정보를 제공하고 있다.

한국어 질의로부터 추출된 문장 구조 정보를 검색 엔진들의 카테고리 정보에 대한 유사도 분석을 통한 정보 유의성 평가는 검색된 정보의 정확도를 매우 개선할 수 있을 것이다.

또한 문형 표준안의 관점에서 일상적 한국어 문장 기술에서 고려될 사항들로 중요성분 생략현상, 자유 어순, 구조적 모호성을 발생시키는 피수식어 선택 문제에 대한 해결 방안을 모색해야 될 것이다.

참고문헌

[1] Ellen M. Voorhees, "Query Expansion Using Lexical Semantic Relations," Proceeding of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp.61-69, 1994.

[2] Hyun-Kyu Kang, Key-Sun Choi "Two-level Document Ranking Using Mutual Information in Natural Language Information Retrieval ," Information Processing & Management, Vol. 33, No.3,pp.289-309,1997.

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

[4] 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.

[5] 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.

[6] 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.

[7] 강현규, 왕지현, 김영섬, 서영훈, “정보 검색에서 질의 형식화를 도와주는 “개념 마법사”의 설계”, 제9회 한글 및 한국어 정보처리 학술대회, pp.23-27,1997.

[8] 김명철, 권오욱, 최기선, 김재군, 김영환, “시소러스와 상호 정보를 이용한 정보 검색 모델”, 1994년도 한국 정보과학회 봄 학술발표논문집 Vol.21,No.1,pp.837-840.

[9] 오종인, 백준호, 최준혁, 이정현, “상호정보량을 이용한 색인어 분류에 의한 웹 정보검색 시스템의 정확도 향상”, 1997년도 한국정보과학회 가을 학술발표논문집 Vol. 24,No.2,pp.201-204,1997.

[10] 정의석, 김기태, 임수종, 차건희, 박재득, 윤보현, 강현규, “정보거래 자동 중개 시스템을 위한 한국어 문형 표준안”, 2000년도 한글 및 한국어 정보처리 학술대회논문집 Vol.12,pp.138-145,2000.