한국어 문장 패턴 기반 개인형 메타 검색 시스템

이덕남° 정혜경 박기선 이용석
전북대학교 컴퓨터과학과
{dnlee°, sysj, kspark}@cypher.chonbuk.ac.kr, {yslee}@moak.chonbuk.ac.kr

A Personalized Meta-Search System based on Korean Sentence Pattern

Deoknam Lee° Heakyung Jeong Kisun Park Yongseok Lee
Dept. of Computer Science, Chonbuk National Univ

요 약

인터넷의 급속한 팽창으로 인해 가용 정보의 양이 폭발적으로 증가하고 있다. 웹 사용자에게 이용 가치가 없는 정보 범람(information overflow)만이 발생한다면 효율적인 정보검색이 되지 못하므로 사용자가 원하는 정보만을 얻을 수 있다면 시간과 미숙한 정보의 검색을 방지 할 수 있다.

본 논문에서는 한국어 질의 생성과 관련하여 웹 사용자의 편의성과 효율성을 고려한 한국어 질의 처리 방법론과 개인형 메타 검색 모델을 제안하고자 한다. 한국어 질의를 기본으로 하여 한국어 문장 패턴 및 개인 정보 평가 구성 요소를 이용한 방법론과 모델을 제안하고자 한다.

1. 서 론

현재 인터넷의 폭발적인 사용 증가로 인해 사용자는 하루에도 셀 수 없을 정도로 많은 양의 정보를 접하게 된다. 인터넷을 이용하는 사용자들은 이러한 수많은 정보 중에서 자신이 실제로 원하는 정보를 구하기가 쉽지 않다. 일반적으로 사용자가 보다 쉽게 원하는 정보를 검색하기 위해서 검색엔진을 이용하는 경우가 대부분이지만 정보의 양이 방대해짐에 따라 검색엔진을 통한 결과물도 사용자들에게는 필요한 정보가 아닐 경우가 많이 발생한다[1].

Yahoo, Excite, Altavista, WebCrawler, Lycos, Google 과 같은 현재 대부분의 인터넷 검색 엔진들은 재현률과 정확률에 대한 문제점들을 외면하고 있다[2]. 몇몇 일반적인 검색엔진들은 질의의 정확성을 개선하는데 메타 검색 엔진들을 이용하려 한다. 예를 들면 MetaCrawler [3], SavvySearch [4], NECI 메타 검색 엔진 [5], Copernic (http://www.copernic.com) 등이 있다.

본 논문에서는 한국어 문장 패턴 기반 개인형 메타 검색 시스템 접근 방법을 제안한다. 한국어 질의어에 나타나는 일정한 질의 유형을 파악하여 일정 패턴형 질의 유형에 따른 주제어를 추출해 냄으로써 정보를 검색하여 사용자로 하여금 찾고자하는 정보가 상위에 랭크되거나 웹 브라우저의 1~2 페이지에 랭크 될 수 있도록 하였다. 또한 사용자들이 구체적으로 그들의 검색 성향을 패턴화된 트리로 나타내는 기법을 제시하였고 다양한 구성요소들을 기반으로 하는 정교한 사용자 선호도 표현 스키마를 구조화하였다.

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 검색 의지를 직접 패턴화 된 트리로 표현하는 것은 편이성 관점에서 매우 제한적이다.

따라서 본 논문에서는 한국어 문장 처리 기법을 이용하여 패턴화 된 트리의 생성을 자동화함으로써 이러한 편이성 문제를 해결하고 동시에 향상된 검색 성과를 거두고자 한다.

본 논문에서는 이러한 한국어 질의 유형에 대해 알아보고 이를 메타 검색 시스템에 적용하여 여러 검색 엔진(야후, 알타비스타, 네이버 등)에서 수집해온 정보를 재순서화 하여 사용자가 원하는 정보가 상위에 랭크되는 개인형 메타 정보 검색 시스템 모델을 제안하고자 한다.

2. 한국어 질의 문장 패턴 유형

2.1 기본문의 개념

기본문은 기본적인 문장패턴에 대한 규정으로서 필수 논항과 서술어로 구성되어 있다[6]. 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 1 항술어, 2 항술어, 3 항술어로 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 1항술어

(문형1)	S → NP1 VP	-1 항술어
	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 고정순서 위반으로 비문이다.

2.2.2 2항술어

(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 고정순서 위반으로 비문이다. 그리고 예문 (4)가 2항술어인 반면, 예문 (5)는 'html과 xml은' NP1이 되는 1 항술어로 본다.

2.2.3 3 항술어

(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 복합 구성이 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 논항 확대이고, (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) $NP \rightarrow (nc|nn)^{*\leq 3}|nb|np$
(규칙 2) $NP' \rightarrow NP$

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. 한국어 문장 패턴 트리와 사용자 선호도 스키마

3.1 형태소 분석과 구문분석 트리 구조

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 문장 패턴을 계층적 개념 트리에 의해 사용자들이 질의어를 색인하고 사용자들의 검색 성향을 나타냈다. 이것을 한국어 문장 패턴 트리 모델(WebWiseViewer(WWV) 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 관계 [그림 1][그림 2]를 표현 할 수 있다.

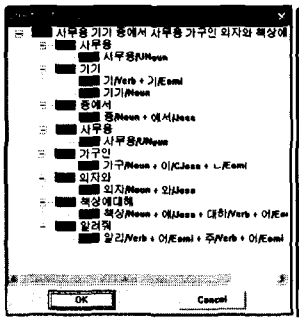

그림 1 형태소분석 결과

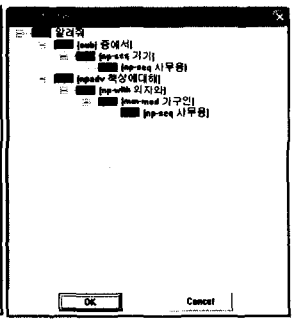

그림 2 구문분석 결과

3.2 사용자 선호도 스키마 개념 모델

다중 결정 척도(multiple decision criteria)에 따른 페이지 순위와 다중 검색 엔진에 의해 제공되는 검색들의 결과를 실험할 것이며 랭킹 처리 방법은 MAU[7]T와 Reportory Grid[8]를 혼합한 형태이며 다음과 같은 6가지 검색 평가 구성요소들을 정의한다.

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

4. 개인형 메타 검색 시스템 구조

이 절에서는 한국어 문장 패턴에 기반을 둔 개인형 메타 검색 시스템의 구조를 나타낸다. [그림 3]은 한국어 문장 패턴에 기반을 둔 개인형 메타 검색 시스템의 전체 구조와 구성요소를 보여주는 중요한 정보의 흐름들을 묘사하고 있다. 한국어 문장 패턴에 기반을 둔 개인형 메타 검색 에이전트 시스템은 8개의 서브시스템과 4개의 중요한 정보 저장소로 구성되어 있다.

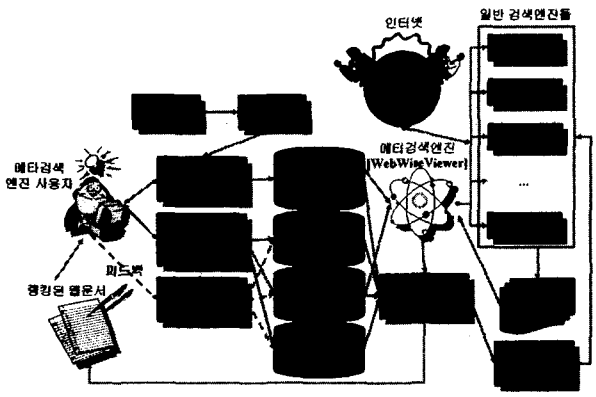

그림 3 개인형 메타 검색 시스템 구조

5. 시스템 구현 및 실험

아래 [그림 4]는 본 논문에서 구현한 한국어 문장 패턴을 기반으로 하는 개인형 메타 검색 시스템의 입력 화면이다. 최종 질의에서는 한국어에 대한 유의어 확장과 질의 변환에서는 한국어에 대응되는 영어 대역어로 확장하여 검색을 할 수 있다[7].

[그림 5]는 한국어 질의에 대해 적절한 불리언 질의를 생성하여 검색이 된 검색 결과의 화면이다.

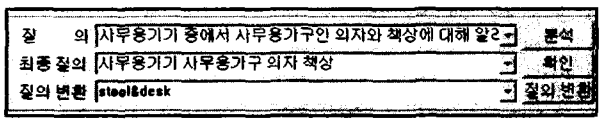

그림 4 개인형 메타 검색 시스템 입력 화면

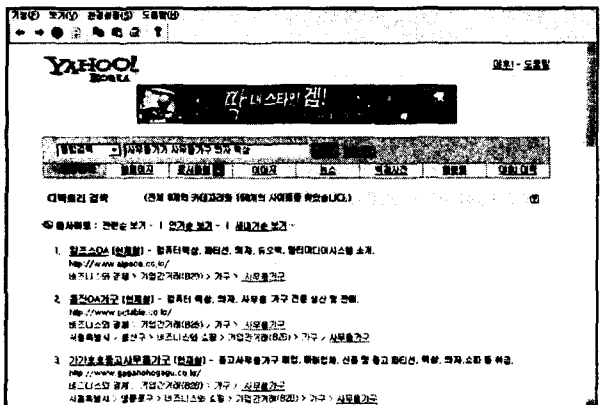

그림 5 개인형 메타 검색 시스템 검색 화면

6. 결 론

본 논문에서는 두 가지 중요하고 보완적인 목표들을 이루기 위해 한국어 문장 패턴 기반 개인형 메타 검색 시스템 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 한국어 질의 유형이 다양해야 할 것이며 문장 패턴 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 한국어 질의의 문장 패턴을 파악한 개념들의 트리화된 구조로서 결정 문제를 표현하고, 구체적인 도메인 명세화 개념에 의해 실제 검색 성향을 사용자들이 표현함으로서 한국어 문장 패턴 기반 트리, 검색 성향 표현 스키마를 제안한다. 웹 페이지 검색을 위해, 잘 알려진 일반 검색 엔진들과 한국어 질의에서 문장 패턴 구조를 파악하여 함께 협력하는 WebWiseViewer(WWV)라고 불리는 한국어 문장 패턴 기반 개인형 메타 검색 시스템 prototype을 설계하였다.

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 패턴 정보를 이러한 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.

본 논문에서는 개별적인 페이지의 새로운 구성요소를 평가하는 사용자의 요구에 따라 현존하는 평가 방법과 함께 고려될 것이다. 이것은 사용자의 변화 욕구와 필요성으로 구성요소의 선택에 의한 개인화를 증가시킬 것이다.

[참고문헌]

[1] 김경만, 이재필, 황수철, 김기태, "사례기반 학습을 이용한 개인형 웹 에이전트의 설계 및 구현", 97년도 한국정보과학회 가을 학술발표논문집(II) 제24권 2호, 1997, pp.97-100.

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

[3] 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.

[4] 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.

[5] 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.

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

[7] 이용석, "한국어 질의어를 수용하는 다국어 정보 검색 엔진 개발", 정보통신부, 산학연 공동기술개발 사업 최종 보고서, 1999.