(제 10회 한글 및 한국어 정보처리 학술대회)

정보검색에서 질의 용어 확장/한정을 위한 자동 질의 용어 정련기의 설계 및 구현

강현수
전북대학교 컴퓨터과학과
전주시 덕진구 덕진동 1가 664-14
우 : 560-756
hs-kang@cs.chonbuk.ac.kr

강현규
전자통신연구원 자연어처리연구부
대전시 유성구 가정동 161
우 : 305-350
hkkang@etri.re.kr

이용석
전북대학교 컴퓨터과학과
yslee@moak.chonbuk.ac.kr

김영섬
전자통신연구원 자연어처리연구부
yskim@etri.re.kr

The Design and Implementation of Automatic Query Term Refiner for Term Expansion/Restriction in Information Retrieval

Hyun-Su, Kang
Department of Computer Science,
Chonbuk University

Hyun-Kyu, Kang
Department of Natural Language Processing
ETRI

Yong-Seok, Lee
Department of Computer Science,
Chonbuk University

Young-Sum, Kim
Department of Natural Language Processing
ETRI

요약

인터넷 정보 검색에서 이용자들이 주로 사용하는 질의는 2-3개의 용어로 이루어진 짧은 질의이다. 또한 동음이의어를 갖는 용어를 사용하기도 한다. 짧은 질의를 처리하는 일반적인 방법은 시소러스[8]나 Wordnet[1]을 이용한 질의 확장이다. 그러나 시소러스나 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. 한편, 사용자의 질의를 주의 깊게 살펴보면, 질의로부터 관련 용어 분류 정보를 추출할 수 있다. 본 논문은 사용자의 질의가 관련 용어 분류 정보에 의해 유기적으로 관계를 가지고 있다는 사실에 기인하여 관련 용어 분류 정보에 따라 자동으로 용어 확장 및 한정을 수행하며 적절한 용어 가중치를 부여하는 자동 질의 용어 정련기를 제안한다. 자동 질의 용어 정련기는 용어의 확장, 한정 및 가중치 부여를 통하여 사용자의 정보 검색 요구를 명확히 하여 검색의 정확성을 향상시킨다.

1 서론

정보의 바다라고 불리는 인터넷에서 사용자가 범하기 쉬운 실수 중 한 가지는 검색 엔진이 지적(intellectual)이라고 생각하는 일이다. 사용자는 검색 엔진이 자신의 질의를 분석하여 자신의 의도에 적합한 검색 결과를 찾아 줄 수 있기를 기대한다. 그래서 사용자는 불과 몇 개의 단어로 이루어진 짧은 질의를 입력하면서 자신이 필요로 하는 정보를 얻고자 한다.

짧은 질의의 질의 확장은 일반적으로 시소러스[8]나 Wornet[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. 그러나 시소러스나 Wordnet과 같은 지식 베이스는 구축하기가 용이하지 않으며, 도메인 종속적인 면과 단어의 희귀(sparseness) 문제를 극복하기 어려운 단점이 있다. 또한, 용어의 동음이의어어로 인하여 검색의 정확성이 떨어지는 문제가 있다.

질의 확장을 위한 방법으로 사용자에게 질의 형식화 도구[6]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 가진다.

또 상호 정보를 정보 검색에 이용하려는 노력에는 [2][8][10]이 있다. [2]는 순위 조정에 상호 정보를 이용하였고, [8]은 시소러스의 자료 희귀를 보완하기 위해 상호 정보를 활용하였다. [10]은 색인어의 의미를 고려하여 색인어 선정 작업에 상호 정보를 이용하였다. [2][8][10]은 상호 정보가 검색 시스템의 정확률을 향상시킬 수 있음을 보여준다.

본 논문에서는 시소러스의 단점을 극복하고 검색 시스템의 정확률 향상을 위하여 코퍼스에서 자동 구축이 가능한 상호 정보 데이터를 사용하여 사용자 질의에서 관련 용어 분류 정보를 추출하고, 검색 의도에 적합하게 자동으로 질의 용어를 확장 및 한정하는 자동 질의 용어 정련기(Automatic Query Term Refiner, AQTR)를 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 자동 질의 용어 정련기의 동기와 의의, 그리고 개념 모델을 알아본다. 3장에서는 자동 질의 용어 정련기의 핵심인 관련 용어 분류와 자동 질의 용어 정련기의 과정을 살펴보고, 4장에서는 시스템 구성을 살펴본다. 5장에서는 자동 질의 용어 정련기를 사용한 검색 예를 살펴보고, 마지막으로 6장에서 결론을 맺는다.

2 자동 질의 용어 정련기

자동 질의 용어 정련기(Automatic Query Term Refiner, AQTR)는 상호 정보를 이용하여 질의로부터 사용자의 관련 용어 분류 정보를 추출하고, 관련 용어 분류 정보에 적합하게 자동으로 질의 용어를 확장 및 한정하며 적절한 가중치를 부여한다.

2.1 동기

인터넷에서 사용자들이 주로 이용하는 자연언어 질의는 2-3개의 짧은 어절로 이루어져 있다. 짧은 질의의 경우, 검색 문서가 없을 수도 있으므로 질의 확장을 한 후 검색을 수행한다. 질의 확장시 동음이의어의 문제가 해결되지 못하면 오히려 검색의 정확성이 떨어진다. 동음이의어로 인하여 사용자들이 의도하지 않은 검색 문서들이 나오는 것은 당연하다.

그러나, 사용자의 질의를 주의깊게 분석해 보면 질의로부터 사용자의 검색 의도, 분류 정보를 찾아낼 수 있다. 자동 질의 용어 정련기는 사용자의 질의가 사용자의 관련 용어 분류 정보에 의해 유기적으로 관계를 가지고 있다는 사실에 기인하여 용어의 확장 및 한정을 수행한다. 용어 확장은 관련 용어 분류 정보에 적합한 용어를 추가함으로 이루어지고, 용어 한정은 사용자가 의도하지 않은 관련 용어 분류 정보를 이용하여 용어를 제외시킴으로 이루어진다. 또한 용어에 대한 적절한 가중치를 부여한다. 용어의 확장 및 한정을 통하여 사용자의 정보 검색 요구를 보다 명확히 할 수 있다.

2.2 의의

상호 정보를 이용하여 질의 용어의 확장 및 한정을 수행하는 자동 질의 용어 정련기의 의의는 다음과 같다. 1)자동으로 용어 확장을 수행함으로 인터넷의 짧은 질의에 대한 대안이 된다. 2)검색 코퍼스에 기반한 상호 정보를 이용함으로 코퍼스 기반 검색이 가능하다. 3)찾고자 하는 사용자의 요구를 명확히 하고 용어의 개념 확장/한정을 통하여 보다 정확한 정보를 찾을 수 있다.

2.3 개념 모델

자동 질의 용어 정련기의 개요는 다음과 같다. 먼저, 질의로부터 검색어를 추출하고, 검색어의 상호 정보 데이터를 분류 알고리즘을 사용하여 분류한다. 분류된 그룹으로부터 사용자의 관련 용어 분류 정보를 찾아낸다. 관련 용어 분류 정보에 적합한 그룹의 경우에는 그룹에서 적절한 용어를 선택하여 첨가하고, 사용자의 검색 의도에 위배되는 경우에는 용어에 검색 X 마크(절대 나오지 말라)를 붙여 용어 제한을 수행한다. 또한 용어에 대한 적절한 가중치를 부여한다.

예를 들면 다음과 같다.
질의 : '지구 공전에 대하여'
1) 검색어 추출 : 지구, 공전
2) 관련 용어 분류 결과

표1 : 지구, 공전의 분류

지구1	행성,월식,지동설,공전, ...
지구2	계절,지구대, ...
공전1	경국대전,속대전,규정....
공전2	민전,경작료,소유,....
공전3	태양,천체,행성,지구,....

3) 용어 확장 및 한정
위 분류로부터 지구1, 지구2, 공전3이 '과학'이라는 같은 도메인을 가지고 있음을 볼 수 있다. 따라서 지구1, 지구2, 공전3에서는 용어를 추출하여 상호정보를 이용한 가중치(여기에서는 0.5를 가정)를 부여하고[용어확장], 공전1, 공전2에서는 용어를 추가하는데 검색 X[용어 한정]를 붙인다. 최종 형식화된 질의는 다음과 같다.

최종질의: 지구:1.0, 행성:0.5, 계절:0.5, 공전:1.0,경국대전:X, 민전:X, 태양 :0.5

위와 같이 용어의 확장 및 한정을 통하여 사용자의 검색의도를 보다 명확히 할 수 있음을 볼 수 있다.

그러나 문제는 표1과 같이 상호 정보 데이터를 분류하여 각각의 분류된 그룹에서 관련 용어 분류 정보를 구하는 것이다. 이를 위하여 본 논문에서는 새로운 분류 알고리즘을 고안하여 상호 정보 데이터를 분류하였고, 분류기[5]를 사용하여 관련 용어 분류 정보를 구하였다. 다음 장에서는 보다 자세하게 관련 용어 분류 알고리즘을 소개한다.

3 관련 용어 분류(LTC)

관련 용어 분류(Local Term Clustering, LTC)는 상호 정보 관계 그래프를 분할하여 2개 이상의 의미있는 서브 그래프로 만든다. 분할된 서브 그래프는 분류기[5]를 이용하여 사용자의 관련 용어 분류 정보를 찾아내는 데 이용된다.

상호 정보 관계 그래프[그림1]는 상호 정보 데이터를 구성하는 각각의 단어들 사이에서 상호 정보의 유무만을 고려하여 그래프로 표현한 것이다. 상호 정보 관계 그래프의 노드는 상호 정보를 구성하는 용어가 되고 에지는 노드들 사이의 상호 정보 관계가 존재함을 의미한다.

3.1 상호 정보

상호 정보[8]는 단어와 단어의 연관성을 정량적으로 나타내기 위하여 사용된다. 코퍼스[11]를 대상으로 추출된 상호 정보 데이터는 표2와 같은 형식을 가지며, 표3은 코퍼스에서 추출한 '동정'에 대한 상호 정보 데이터 예이다.

표2 : 상호정보 데이터의 형식

상호정보이름	수	용어	값

첫번째 필드는 상호 정보 데이터의 이름을 나타낸다. 두번째 필드의 수는 첫번째 필드의 용어와 상호 정보량이 존재하는 용어의 수를 나타낸다. 이 수만큼 용어와 상호 정보 값이 반복적으로 나오게 된다.

표3 : '동정'의 상호 정보 데이터

동정	55	겉깃	2.034729	덧꾸미	2.034729	목둘레선	
2.034729	숨뜨기	2.034729	오스틴	2.034729			
태서문예신보	2.034729	편모방	2.034729	청빈			
1.858638	타국	1.733699	수녀	1.636789	해면동물		
1.636789	조붓	1.557608	메데이아	1.432669	순결		
1.189631	독신	1.189631	공감	1.189631	간첩		
1.131639	섶	1.080487	복종	1.057006	서약		
0.993337	맹세	0.920786	편모	0.920786	겉겹질		
0.888601	두루마기	0.804280	고름	0.755976	사제		
0.712510	한복	0.702291	복음	0.654518	수도원		
0.619756	신부	0.611483	사실주의	0.587571			
저고리	0.579884	바늘땀	0.529579	심리적			
0.529579	소박	0.516215	이성	0.503250	차림		
0.490661	남성	0.484501	운명	0.478427	내정		
0.454946	해석	0.454946	오리	0.443665	헝겊		
0.411480	성적	0.362631	문단	0.327159	관찰		
0.306375	카톨릭교	0.231955	여성	0.211908	흰		
0.183471	해외	0.142635	소매	0.120915	일상		
0.100231	소개	0.070941	묘사	0.057006	번역		
0.001305							

표3은 '동정'이라는 용어가 55개의 용어와 상호 정보의 관계가 있음을 보여준다.

표 3 '동정'의 상호 정보 데이터를 그래프로 표현한 것이 그림1 '동정'의 상호정보 관계 그래프이다.

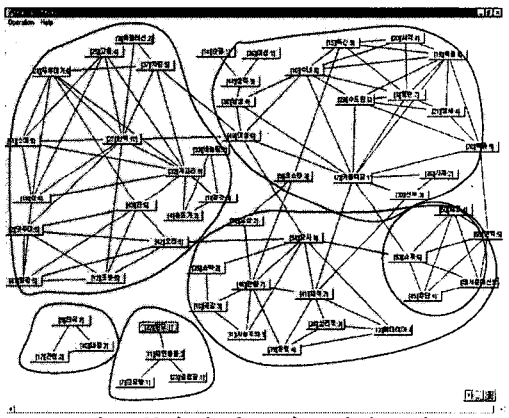

그림1 : '동정'의 상호 정보 관계 그래프

표3은 위 그림1 '동정'의 상호 정보 관계 그래프에서 보듯 5개의 개념적인 서브 그래프로 분할되어진다. 5개의 개념적인 서브 그래프와 중심 용어는 표4와 같다.

표4: 상호정보 데이터 동정의 개념적인 분류

그룹	중심 용어
A : 동정1	한복
B : 동정2	카톨릭교
C : 동정3	간첩
D : 동정4	해면동물
E : 동정5	묘사

실제 '동정'의 사전적 분류[부록]와 비교해 보면 약간의 차이가 있다. 부록에서 보여주는 것처럼

사전적 분류의 동정2, 동정3, 동정4는 표4에는 존재하지 않는다. 이러한 차이는 코퍼스에서 사전적으로 분류된 모든 의미가 사용되지 않기 때문에 일어난다. 그러나, 표4의 D,E와 같은 그룹은 사전적 분류에는 포함되어 있지 않다. 이유는 상호 정보가 용어와 용어의 정량적인 관계를 나타내기 때문이다.

3.2 LTC 알고리즘

관련 용어 분류는 상호 정보 데이터에 의하여 표현된 상호 정보 관련 그래프를 2개 이상의 서브 그래프로 분할한다. 이미 2개 이상의 서브 그래프로 분할된 상호 정보 관련 그래프일지라도 관련 용어 분류 알고리즘을 적용하여 미세한 분류를 행한다.

관련 용어 분류에 사용되는 알고리즘은 다음과 같다.

n : 상호 정보 데이터 수
상호 정보 관계 그래프의 노드 수

단계1) 상호 정보 관계 그래프 표현 : n(상호 정보 데이터 수)차원 정방 행렬 안에 0(에지없음),1(에지존재)을 사용하여 상호 정보 관계 그래프를 표현한다.

단계2) 길이n의 배열 안에 용어와 용어의 총 에지 수를 구하고, 총 에지 수에 따라 내림차순으로 정렬한다. 각 노드의 방문 정보는 거짓으로 초기화되며, 구해진 총 에지 수는 노드의 링크 수라는 이름으로 참조된다.

단계3) 그래프 방문은 단계 2의 배열에서 에지의 내림차순으로 순차적으로 이루어진다. 분할 그래프 G는 방문하지 않은 가중치가 가장 큰 노드 N을 중심으로 가중치의 내림차순으로 연결된 방문하지 않은 모든 노드가 포함된다. 이 노드들은 스택에 push되어 다음 방문을 위해 준비되며 방문 정보는 참이 된다. 자신보다 하나라도 큰 가중치를 가진 노드를 만나면 방문은 중단된다.

단계4) 방문이 중단되면 스택에서 모든 노드를 pop할 때까지 단계3을 반복하며 그래프 G를 확장한다.

단계5) 모든 노드를 방문할 때까지 다음 분할 그래프 G'를 구하기 위해 단계 3을 반복한다.

그래프 방문 과정 : 그림2의 괄호 안의 숫자는 링크 수를 나타내며, 방문 과정은 다음과 같다. 먼저, 가장 큰 링크 수를 가지는 카톨릭교(11)에서 여성(6)을 방문한다. 그러나 여성(6)은 자신보다 큰 링크 수를 가지는 한복(10)과 연결되어 있기 때문에 방문은 중단되고 분할 그래프 G = {카톨릭교, 여성}가 구해진다. 다음 그래프 G'는 방문하지 않은 노드 중 가장 큰 링크 수를 가지는 한복(10)을 시작점으로 여성(6)은 이미 방문한 노드임으로 차림(4)에서 방문이 중단된다. 따라서 분할 그래프 G'={ 한복, 차림 }이 구해진다.

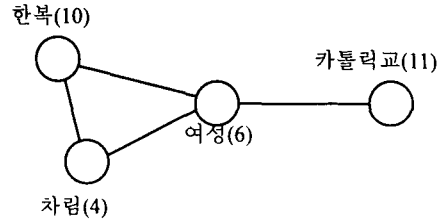

그림2 : 그래프 방문 과정

3.3 LTC 결과

표5는, 표3 '동정'의 상호 정보 데이터를 위 알고리즘을 이용하여 LTC를 적용했을 때 얻어진 결과이다.

표5 동정의 LTC 적용 결과

그룹	단어 리스트
A	카톨릭교 소개 번역 해외 문단 태서문예신보 여성 해석 운명 심리적 남성 성적 이성 순결 신부 수도원 복음 사제 복종 맹세 서약 독신 수녀 청빈
B	한복 소매 차림 저고리 흰 헝겊 숨뜨기 바늘땀 겉깃 고름 두루마기 섶 조붓 목둘레선 덧꾸미
C	묘사 관찰 일상 공감 오스틴 오리 소박 사실주의 메데이아
D	해면동물 겉껍질 편모 편모방
E	간첩 내정 타국

이 결과는 표4의 개념적인 분류와 결과가 일치함을 보여준다. 차이가 있다면, 그림 1에서 F의 영역이 위 표에서는 그룹 A 카톨릭교(밑줄 단어들)에 포함되어 있음을 볼 수 있다(본래 F영역은 묘사 그룹에 포함되어 있었다.) 이는 관련 용어 분류 알고리즘이 노드의 링크 수를 사용하여 포함여부를 결정하는데, '카톨릭교'의 링크 수가 '묘사'의 링크 수보다 상대적으로 크기 때문에 F영역이 모두 그룹 A 카톨릭교에 배정되었다.

참고로 '베이컨'이라는 용어의 상호 정보 데이터를 관련 용어 분류에 적용하면 표6과 같다.

표6 : '베이컨'의 LTC 적용 결과

그룹	단어 리스트
A	철학 계승 관직 직위 박탈 뇌물 대법관 철학자 문인 셰익스피어 런던 변호사 판사 법률학 스펜서 엘리자베스1세 중요시 사상가 창시자 논문 귀납법 적용 수필 감옥 유물론 홉스 경험론 학풍 신기관
B	햄 반찬 소시지 가공품 돼지 관찰 랜드레이스종 옆구리살 훈연 돼지고기 훈제품 뼈햄

3.4 카테고리 설정

카테고리를 설정하기 위하여 분류기[5]를 사용하였다.

표7과 8은 '동정'과 '베이컨'의 카테고리 설정 결과이다. 분류기를 이용할 때 단어가 3개 이하인 그룹은 분류의 정확성을 고려하여 버리고, 카테고리를 설정하였다.

표7 : '동정'의 카테고리 설정

그룹	중심단어	분류코드	분야
A	카톨릭교	203	크리스트교카톨릭교
B	한복	611	가정가사
C	묘사	102	논리학
D	해면동물	501	동물일반

표8 : '베이컨'의 카테고리 설정

그룹	중심단어	분류코드	분야
A	철학	199	철학관련인명
B	햄	612	식품영양

3.5 용어 확장/한정

용어 확장에 추가되는 용어의 수는 실험을 통하여 적정한 임계치(threshhold)를 선정해야 한다. 그러나 본 논문에서는 실험 상의 편의를 위하여 추가되는 용어의 수를 1로 제한하였다.

질의 '동정은 옷에서 어떻게 쓰이나'에 대한 용어 확장 및 한정의 절차는 다음과 같다.

1) 검색어 추출 : 옷, 동정
2) 관련 용어 분류 및 카테고리 설정
   관련 용어 분류를 거쳐 카테고리를 설정한 결과는 표7, 표9와 같다.

표9 : '옷'의 카테고리 설정

그룹	중심단어	분류코드	분야
A	모양	611	가정가사
B	관리	611	가정가사

3) 공통 카테고리 추출
   옷, 동정, 쓰임새에서 공통 카테고리 정보를 추출하면 611(가정가사)로 일치한다. 만약 소분류 코드가 일치하지 않으면 대분류를 적용하여 공통 카테고리를 추출한다.
4) 용어 확장 및 한정
   공통 카테고리 611(가정가사)과 같은 분류코드를 가지는 그룹에서는 용어를 추가하여 용어 확장을 수행하고, 다른 그룹에서는 용어를 추가하되 X(절대 나오지 마라) 마크를 붙여 용어를 첨가하여 용어 한정을 수행한다. 용어 추가시 중심 단어를 추가하였으나 1자짜리 단어는 모호성 때문에 배제하였다. 1자짜리 단어가 중심 단어인 경우는 다음 단어를 추가하였다. 용어 확장 및 한정의 결과는 다음과 같다.

   옷A : 모양
   옷B : 관리
   동정A : 카톨릭교X
   동정B : 한복
   동정C : 묘사X
   동정D : 해면동물X

참고로, 질의 '베이컨과 소시지에 대하여'에 대한 용어 확장 및 한정은 다음과 같다.

1) 검색어 추출 : 베이컨 소시지
2) 관련 용어 분류 및 카테고리 설정
   추출된 검색어의 카테고리 설정 결과는 표8과 표11과 같다.

표11 : '소시지'의 카테고리 설정

그룹	중심단어	분류코드	분야
A	들	612	식품영양
B	만	612	식품영양

3) 공통 카테고리 추출
   공통 카테고리를 추출하면 612(식품영양)을 얻을 수 있다.
4) 용어 확장 및 한정
   '베이컨B'와 '소시지'의 중심 용어가 1자짜리 단어이므로 무시하고 다음 단어를 중심 용어로 선택하여 얻은 결과는 다음과 같다.
   베이컨A : 철학X
   베이컨B : 반찬
   소시지A : 가지
   소시지B : 가공

3.6 가중치 부여

가중치 부여 단계에서는 용어 확장으로 첨가된 용어에 적절한 가중치를 부여한다. 본 논문에서는 개념적으로 가중치를 부여하였다.

개념적으로 가중치를 부여하는 기준은 다음과 같다. 질의에서 추출된 검색어는 기본적으로 가중치 1.0을 할당하고 공통 카테고리가 존재하는 경우, 소분류 코드가 일치하는 경우는 1.0보다 작으면서 대분류 코드의 가중치보다 크게 주고, 대분류 코드가 일치하면 가장 작은 가중치를 할당한다. 가중치는 실험을 통하여 적정한 임계치(threshhold)를 설정해야 할 것이다. 그러나 본 논문에서는 실험의 편의를 위하여 질의에서 추출된 검색어는 1.0을 할당하였고, 소분류 코드가 일치하는 경우에는 0.5를, 대분류 코드가 일치하는 경우에는 0.3를 할당하였다.

가중치 부여 결과

1) 질의 : 동정은 옷에서 어떻게 쓰이나
옷 :1.0, 모양 :0.5
동정 :1.0, 카톨릭교X, 한복 :0.5, 묘사X,
해면동물X
2) 질의 : 베이컨과 소시지에 대하여
베이컨 :1.0, 철학X, 반찬 :0.5
소시지 :1.0, 가지 :0.5, 가공 :0.5

4. 시스템 구성

4.1 전체 구성

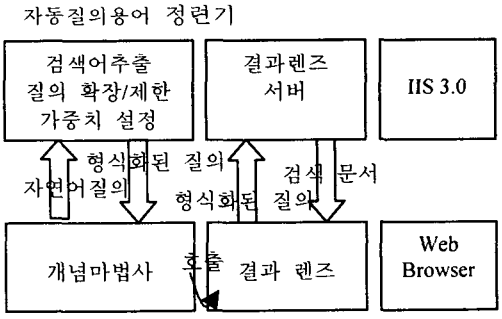

그림3 : 전체 시스템 구성도

자동 질의 용어 정련기는 그림3에서 보는 것처럼 개념 마법사[6]와 결과렌즈[7]를 근간으로 하고 있다. 개념 마법사는 질의 형식화 도구로 사용된다. 개념 마법사를 통해 사용자가 접하는 초기 형식화된 질의는 자동 질의 용어 정련기에서 1차적으로 용어의 확장과 한정, 가중치 부여가 이루어진 검색어들로 이루어져 있다. 사용자는 개념 마법사를 이용하여 자동 질의 용어 정련기의 질의 확장의 오류 수정은 물론 검색어를 보다 세밀하게 정련(refine) 할 수 있다.

4.2 AQTR 구성

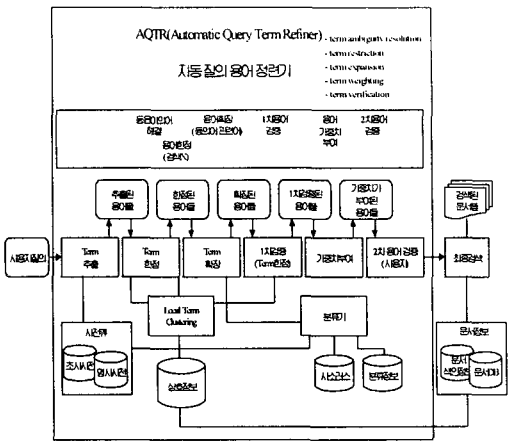

그림 4 : 자동 질의 용어 정련기의 구성도

그림4는 자동 질의 용어 정련기의 구성도이며, 사용 절차는 다음과 같다. 1)먼저 사용자의 질의로부터 검색어들을 추출한다. 2)추출된 검색어들의 상호 정보 데이터를 이용하여 관련 용어 분류(Local Term Clustering) 작업을 실시한다. 3)사용자의 관련 용어 정보와 일치한 분류 그룹에서는 용어 추가(용어확장)가 이루어지는 반면, 그렇지 않은 그룹에서는 용어 한정(검색X)이 이루어진다. 4)검색어에 가중치를 부여한다. 5)자동 질의 용어 정련기에 의하여 최종 형식화된 질의는 개념마법사에게 전달되어지고, 사용자는 자동 질의 용어 정련기의 오류를 수정하거나 보다 세밀하게 정련한 후 검색을 시작한다.

만약 자동질의용어 정련기에서 사용자의 질의로부터 공통 카테고리를 찾아내지 못했을 경우는 질의에서 추출한 검색어들만 개념마법사에게 전달한다.

5. 검색 예

5.1 일반 검색

동음이음어가 있을 경우 검색의 정확성이 떨어질 수 밖에 없음을 그림5의 개념 관계 그래프는 보여준다.

(사용자의 요구가 굵은 실선으로 표시되어 있고, 가는 실선은 동음이의어의 모호성을 나타낸다.) '동정'이라는 단어가 5가지 개념으로 쓰이기 때문에 중의성을 배제하지 않는 일반검색 시스템에의 경우, '동정은 옷에서 어떻게 쓰이나'라는 질의는 실제 '옷', '동정' 만을 가지고 질의를 한다고 할 지라도 '옷', '동정', '카톨릭', '한복', '묘사', '해면동물', '간첩'의 검색어를 가지고 검색을 하는 것과 같은 효과를 가진다.

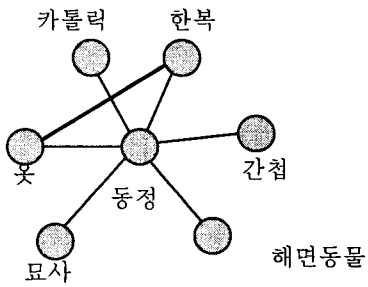

그림5 : '동정은 옷에서 어떻게 쓰이나'의 개념 관계 그래프

실제 '동정은 옷에서 어떻게 쓰이나'라는 질의에서 검색어만을 추출하여 검색을 실행한 결과 상위 10개 문서에서 사용자가 요구한 문서는 2개에 불과하였다. 표12는 상위 10개의 문서와 관련성을 보여준다.

표12 : 일반 검색의 경우

순위	문서 제목	관련성
1	동정2	X
2	동정1	O
3	밀레	X
4	메테이아	X
5	간첩	X
6	해면동물	X
7	태서문예신보	X
8	숨뜨기	O
9	수도원	X
10	수녀	X

5.2 AQTR 검색

자동 질의 용어 정련기는 사용자 질의에서 사용자의 의도, 분류 정보를 추출한 뒤에 검색을 하기 때문에 사용자가 요구를 명확히 할 수 있다. 그림5에서 보다시피 굵은 실선으로 표현된 것이 자동 질의 용어 정련기의 개념적인 관계 그래프이다.

그림6과 그림7은 질의 '동정은 옷에서 어떻게 쓰이나'에 대한 AQTR의 실행 결과와 용어 확장, 한정 및 개념적 가중치를 부여하여 얻은 검색 결과이다.

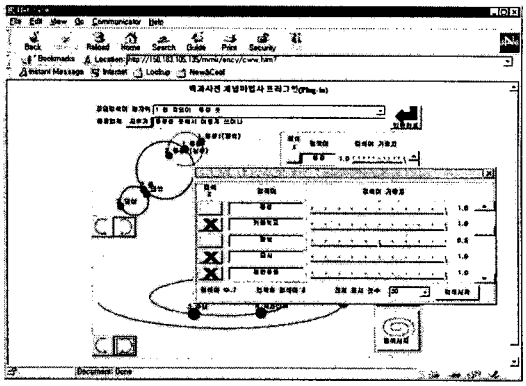

그림 6 : AQTR의 실행 결과

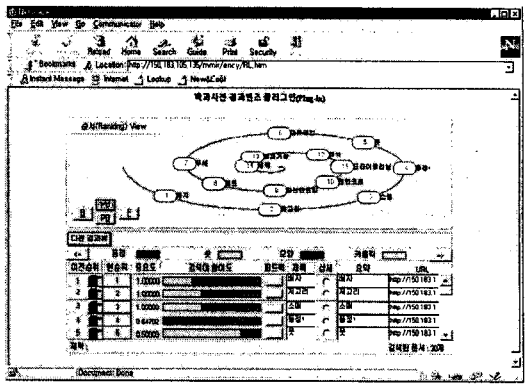

그림 7 : 용어의 확장, 한정 및 개념적 가중치를 부여하여 얻은 검색 결과

표13은 그림 7 용어 확장, 한정 및 개념적 가중치를 부여하여 얻은 검색 결과이다.

표13 : 용어 확장, 한정 및 개념적 가중치를 부여한 AQTR의 검색 결과

순위	문서 제목	관련성
1	배자	O
2	저고리	O
3	소매	O
4	동정1	O
5	옷	O
6	의류공업	O

7	푸새	O
8	벨트	O
9	맞선단트임	O
10	레인코트	O

표12 일반 검색의 결과와 표13 용어의 확장, 한정 및 개념적 가중치를 부여한 AQTR의 검색 결과를 비교해 보면 다음과 같다. 표12의 검색 결과를 살펴보면, '동정'의 모호성이 해결되지 않아 검색 결과가 다양한 카테고리를 가짐을 볼 수 있다. 그러나 표13의 경우 '옷'에 의하여 '동정'의 의미가 '가정가사'의 카테고리로 한정되어 문서 결과 전체가 '가정가사'의 카테고리를 가진다.

6. 결론

본 논문에서는 질의로부터 사용자의 관련 용어 분류 정보를 추출할 수 있다는 사실에 기인하여, 질의 확장, 한정 및 가중치 부여를 자동으로 수행하는 자동 질의 용어 정련기를 제안하였다. 자동 질의 용어 정련기는 상호정보를 이용하여 질의 확장을 하기 때문에 인터넷의 짧은 질의에 대한 대안이 대며, 특히 동음이의어의 경우 용어의 개념 확장, 한정 및 가중치 부여를 통하여 찾고자 하는 사용자의 요구를 명확히 할 수 있음을 보였다. 그러나 보다 많은 질의를 통한 실험 및 평가 분석이 요구되며, 용어 가중치 부여시 경험적인 방법 이외에 적절한 가중치 부여 알고리즘을 필요로 한다.

참고문헌

[1] Ellen M. Voorhees, "Query Expansion Using Lexical-Semantic Relations," Proceedings 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] Mark Magennis and Cornelis J. van Rijsbergen, "The potential and actual effectiveness of interactive quer, expansion," Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.324-331, 1997

[4] Peter G. Anick and Shivakuar Vaithyanathan, "Exploiting Clustering and Phrases for Context-Based Informational Retrieval," Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.314-323, 1997

[5] 강원석, 강현규, 김영섬, "개념 기반 문서 분류기 TAXON의 설계 및 구현", 1997년도 한국정보과학회 추계 학술발표 논문집(B), pp.197-200, 1997.

[6] 강현규, 왕지현, 김영섬, 서영훈, "정보 검색에서 질의 형식화를 도와주는 "개념마법사"의 설계", 제9회 한글 및 한국어 정보처리 학술대회, pp.23-27, 1997.

[7] 강현수, 강현규, 김영섬, 이용석, "대량의 정보 검색 결과를 위한 "결과 렌즈"의 설계 및 구현", 1998년도 한국정보과학회 봄 학술발표논문집(B) Vol. 25, No. 1, pp.449-451, 1998.

[8] 김명철, 권오욱, 최기선, 김재군, 김영환, "시소러스와 상호정보를 이용한 정보 검색 모델", 1994년도 한국정보과학회 봄 학술발표논문집 Vol. 21, No.1, pp.837-840

[9] 동아출판사, 동아 새국어사전, pp.607-608, 1995

[10] 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.

[11] 한국전자통신연구원, "ETRIKEMONG SET", 자연어처리연구실, 한국전자통신연구원, 1997

부록1

'동정'의 사전[9]적 분류

1. 동정(명) : 한복에서, 저고리 깃 위에 조붓하게 덧대는 흰 헝겊 오리.
2. 동정(同定)(명)(하|자) 생물의 분류학상의 소속을 정하는 일.
3. 동정(同情)(명)(하|타) 남의 불행이나 슬픔 따위를 자기 일처럼 생각하여 가슴 아파하고 위로함.
4. 동정(東征)(명) 동방을 정벌함.
5. 동정(動靜)(명) (어떤 행동이나 상황 등이) 전개되거나 변화되어 가는 낌새나 상태.
6. 동정(童貞)(명)1)이성과 아직 성적 관계를 가진 일이 없는 사람, 또는 그러한 상태. 2)카톨릭에서, '수도자'를 이르는 말. 3)동정남(童貞男).