Journal of Korea Multimedia Society Vol. 18, No. 1, January 2015(pp. 52-64)
http://dx.doi.org/10.9717/kmms.2015.18.1.052

엔터테인먼트 데이터를 위한 자연어 검색시스템

김 정 인†

A Natural Language Retrieval System for Entertainment Data

Jung-In, Kim†

ABSTRACT

Recently, as the quality of life has been improving, search items in the area of entertainment represent an increasing share of the total usage of Internet portal sites. Information retrieval in the entertainment area is mainly depending on keywords that users are inputting, and the results of information retrieval are the contents that contain those keywords. In this paper, we propose a search method that takes natural language inputs and retrieves the database pertaining to entertainment. The main components of our study are the simple Korean morphological analyzer using case particle information, predicate-oriented token generation, standardized pattern generation coherent to tokens, and automatic generation of the corresponding SQL queries. We also propose an efficient retrieval system that searches the most relevant results from the database in terms of natural language querying, especially in the restricted domain of music, and shows the effectiveness of our system.

Key words: Natural Language, Query, Retrieval System, Entertainment, pattern

1. 서 론

최근 한류를 기반으로 엔터테인먼트 분야가 급격히 성장함에 따라서 관련 콘텐츠에 대한 검색도 크게 증가하고 있다. 국내 포털 검색 서비스 업체들의 인기검색어 대부분은 예능, 음악, 영화, 드라마 등 엔터테인먼트 콘텐츠들이 차지하고 있는 실정이다.

국내 검색사이트를 통한 엔터테인먼트 관련 검색어의 결과는 대부분 관련 콘텐츠 축약 정보, 뉴스기사, 웹 게시물 등 키워드 중심의 관련결과를 찾아준다. 하지만, 현재 시스템에서 우리는 “아이유와 혈액형이 같은 연예인”이라던지 “수지보다 나이가 어린 연예인” 혹은 “JYP 소속의 연예인들”, “영화 역린에 나온 배우들”, “김경호가 부른 노래들”과 같은 질의에 대한 정확한 답변을 기대할 수 없다. 한류 중심의 엔터테인먼트와 관련된 데이터베이스를 구축한 후, 다양한 자연어 검색에 대해 정확한 결과를 제공해 줄 수 있다면 많은 사람들에게 도움이 될 것이고 특히 한류 확산에도 크게 이바지할 것으로 기대가 된다.

한국어 자연어 검색은 엠파스 검색엔진 등에 탑재되어 한동안 각광을 받다가 복잡한 언어처리과정에 비해 검색결과에 대한 만족도가 높지 않다는 이유로 점차 도태된 실정이다. 하지만 최근에 아이폰의 음성인식 자연어처리모듈인 시리를 필두로 간단한 번역이 행해지는 앱까지 다양하게 출시되면서 음성인식과 연계되는 자연어 처리기술이 다시 화두에 오르고 있다. 간소화된 자연어처리 기술이 개발된다면 많은 곳에서 활용할 수 있는 여지가 있다.

본 연구에서는 한류와 관련되는 엔터테인먼트 데이터를 다양한 웹사이트에서 모아오는 웹크롤링시

※ Corresponding Author : Jung-In Kim, Address: (608-711) Sinsun-ro 428, Nam-gu, Busan, Korea, TEL : +82-51-629-1174, FAX : +82-51-629-1169, E-mail : ji-kim@tu.ac.kr
Receipt date : Sep. 4, 2014, Revision date : Dec. 10, 2014
Approval date : Dec. 16, 2014
† Dept. of Computer Engineering, Tongmyong University
※ This Research was supported by the Tongmyong University Research Grants 2012.(2012A004)

스템을 이용하여 엔터테인먼트 데이터베이스를 구축하고 엔터테인먼트와 관련된 자연어 검색 문장을 수집하여 정형화된 패턴을 구축한다. 그 후, 자연어 질의에 대해 간이 형태소분석과 구문분석 후, 패턴 매칭을 시행하고 가장 가까운 패턴의 SQL 질의어로 생성하여 데이터베이스를 검색하는 작업을 통하여 자연어 질의에 대해 답변할 수 있는 방법을 제안한다.

2. 관련 연구

기존의 엔터테인먼트 관련정보 검색은 포털사이트에서 많이 이루어지고 있다. 네이버(NAVER), 다음(DAUM), 네이트(NATE) 그리고 구글(Google)은 엔터테인먼트 관련 정보 검색이 가능하며 사용자는 쉽게 관련된 콘텐츠의 정보를 구할 수 있다. 엔터테인먼트 정보는 분야(영화, 음악, 인물 등)에 따라 각각의 다른 서비스로 제공한다. 그러나 서비스간의 연계가 없어서 인물과 콘텐츠, 콘텐츠와 콘텐츠 간의 연관성을 알기 어렵다. 네이트의 경우, 인물 검색을 중심으로 다른 업체와 차별화된 서비스를 제공하고 있다. 특정 인물의 개인 정보를 최근이슈, 프로필, 이야기, 주요작품, 뮤직, 인물관계도로 분류하여 제공하고 타임라인을 두어, 그 들의 시간적 이슈를 살펴볼 수 있도록 하였다. Fig. 1은 네이트와 네이버의 연예인 “보아”에 대한 검색결과를 나타낸 것이다.

데스티(Desti)는 자연어 기반 검색과 인공지능 기술을 접목한 여행정보 검색용 애플리케이션이다. 시리(Siri)가 처음 개발되었던 비영리 연구기관인 SRI International에서 2011년부터 개발을 시작했으며, 현재 아이패드 앱이 출시된 상황이다. “뉴욕 맨하탄에서 이틀동안 머물 예정인데 가볼만한 곳들은?” 혹은 “미국 대륙횡단을 하려고 하는데 다양한 국립공원을 경험해 볼 수 있는 가장 좋은 경로는?” 등의 여행과 관련된 질의에 최적의 정보를 사용자의 기호에 맞게 제공하는 것이 목표이다. 단순히 음성을 문자로 변환하여 검색하는 것이 아니고 사용자의 기호를 이해해서 계획 중인 여행에 도움이 될 검색결과만을 제공하겠다는 것이며, 당분간은 미국에서만 서비스를 제공할 계획이다[1].

SNS와 스마트폰 활용기술이 발달함에 따라 음성처리와 같은 대화체 기반의 자연어 처리도 요구되고 있다. 기존의 한국어 형태소 해석 기법들은 문어체 기반의 정형화된 문장을 위주로 개발되어 왔으므로 비문법적인 요소가 포함된 대화체 문장의 해석에는 적합하지 않았다. 정성훈과 황도삼은 대화체 문장의 특징을 분석하고 음절 단위의 형태소 해석 기법에 기반한 대화체 문장용 형태소 해석 기법을 제시하였다[2]. 그리고, 자연어로 된 한국어 질의분석을 위하여 다양한 논문이 보고되었다. 우선, 구문구조를 파악하는 경우가 있다. 한국어 질의에 대해 의문사를 이용하여 사용자의 질의를 6개의 상위 범주 질의 유형으로 분류하고, 술어 유형 정보와 구문 구조 정보를 이용하여 초점 단어를 추출한다. 추출된 초점 단어의 질의 유형 자질을 이용하여 53개의 하위 범주 질의 유형으로 세분화하여 분류한다. 그리고 술어기반 질의분석을 위해 술어로 끝나는 질의와 술어가

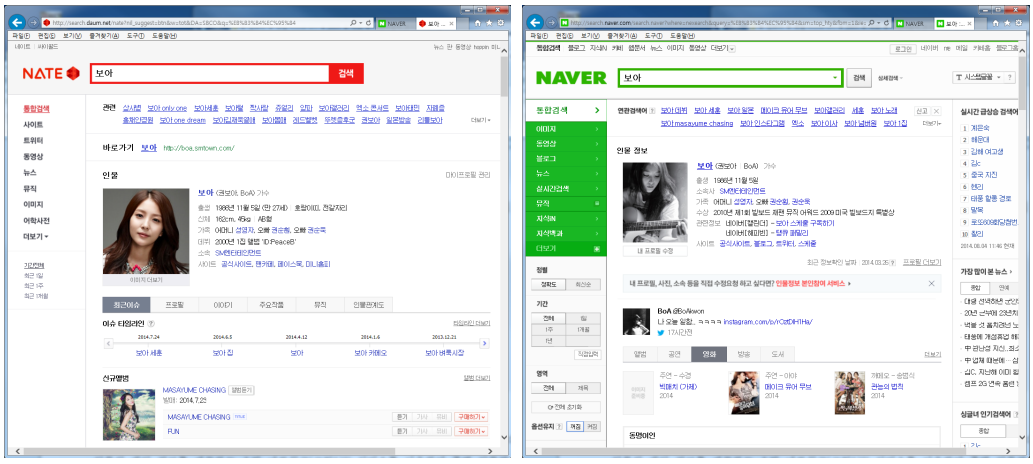


Fig. 1. Search Results of “Boa” on Nate and Naver.

생략된 질의로 나누어 초점단어를 추출한다. 구문구조 정보는 명사구의 구조와 그 구조에서의 초점 단어가 될 수 있는 명사들의 경향을 나타낸다. 초점 단어로 추출된 명사구나 술어가 생략된 질의가 구문구조 정보에 있는 명사구 구조와 일치할 경우, 초점 단어가 될 수 있는 경향이 가장 큰 명사를 초점 단어로 추출하는 방법을 제안하였다[3]. 또한 신승훈, 서영훈은 술어 유형정보를 이용하여 대분류 수준의 정답 유형으로 질의분석을 수행하고 구문 구조 정보를 이용하여 중요 키워드와 일반 키워드를 추출하고 정답 유형 자질 명사를 이용하여 세부 정답유형을 결정하는 방법을 제안하였다[4].

특허검색에만 최적화된 검색시스템도 연구되었다. 특허를 검색하는 기존 방법은 불리언 모델을 기반으로 단어의 존재 여부만을 파악하는 방식이 주로 사용되었지만 검색 결과에 노이즈 데이터가 너무 많이 포함되어 있어서 정확한 특허 검색에 오랜 시간을 허비하게 만들어서, 결국 전문 검색사들이 수동으로 찾아주고 있는 실정이었다. 이우기, 송종수, 강민구는 기존의 일반적 문서검색과 특허검색과의 차이점을 밝히고, 기존 특허검색의 한계성을 분석하였다. 나아가 특허검색에 특화된 효과적 방법론을 제안하여 검색 질의어가 각 특허 문서 내에서 차지하는 중요도와 각 문서 내에서 질의어 사이의 관계성을 파악하고 이에 대한 랭킹을 정하여 질의어와 관계성이 높은 특허가 상위에 랭크하고 노이즈 데이터를 하위에 랭크 시킴으로써 검색결과의 효율을 높이는 방법을 제안하였다[5].

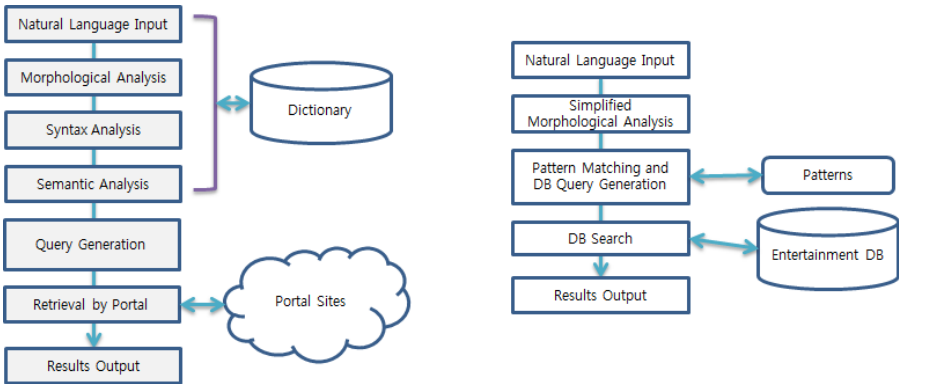
질의문을 심층 분석하는 연구도 진행되었다. 일반적인 질의응답 시스템들은 사용자가 입력한 자연어 질의의 의미를 분석하지 않기 때문에 정확한 대답을 제공하는 것이 어렵다. 질의문 심층 분석은 의미자질 추출 문법과 자연어 질의 특성을 이용하여 사용자의 질의를 의미적으로 분석하고, 의미자질들을 추출한다. 의미자질 추출 문법과 자연어 질의 특성은 사용자 질의의 의미와 구문 구조를 반영하기 위해 의미자질과 형식형태소로 표현된다. 이들은 웹에서 추출한 세부 정답 유형이 '인물'인 100개의 질의에 대한 실험을 통해, 비교적 짧지만 사용자의 질의 의도를 충분히 표현하고 있는 자연어 질의에 대해 질의문 심층 분석을 수행함으로써 사용자의 질의 의도를 분석하고, 의미자질들을 추출할 수 있음을 보였다[6].

고성능의 질의응답 시스템을 구현하기 위하여 사용자의 질의 의도를 파악할 수 있는 질의 유형 분류기에 대한 연구도 진행되었다. 우선, 사용자 질의에 포함된 어휘, 품사, 의미표지와 같은 다양한 정보를 이용하여 사용자 질의로부터 자질들을 추출한다. 다량의 자질들 중에서 유용한 것들만을 선택하기 위해서 카이 제곱 통계량을 이용한다. 추출된 자질들은 벡터 공간 모델로 표현되고, 문서 범주화 기법 중 하나인 지지 벡터 기계는 이 정보들을 이용하여 질의 유형을 분류한다. 질의 유형 분류기를 통계적 방법으로 구축함으로써 lexico-syntactic 패턴과 같은 규칙을 기술하는 수작업을 배제할 수 있었다[7]. 그 외 한국어 질의 분석과 관련한 연구로 의문대명사의 패턴을 정의하여 한국어 질의어를 분석하는 시도[8]와 자연어 질의를 객체지향 데이터베이스의 질의어인 OQL로 변환시키는 시도[9] 등이 있었다. 김기철은 질의어로부터 사용자의 일반적인 요구사항들을 에러없이 뽑아내기 위하여 의문대명사를 고려하였다. 한국어로 질의어를 작성할 때 제한된 의문대명사의 사용을 허용하기 위해 4개의 대명사 패턴을 정의하고, 질의어에서 의문대명사의 패턴을 결정하면 의문대명사가 가진 의미와 정보를 가지는 키워드로 대체하여 질의 의도를 명확하게 하는 것이다. 채진석은 객체지향 데이터베이스를 위한 한국어 질의 시스템을 설계 및 구현하였는데, 한국어 질의를 OQL명령문으로 변환하기 위해 프레임기반 질의 분해 기법을 사용하였다. 사용자에 의해 입력된 한국어 질의는 한국어 파서에 의해 파스 트리로 변환되고, 이 파스 트리는 다시 질의구로 분해된다. 이렇게 분해된 질의구는 질의 프레임으로 변환되며, 이 질의 프레임으로부터 목표 OQL명령문이 생성된다.

또한 시맨틱 웹 구현의 중요한 수단인 온톨로지는 검색, 추론, 지식표현 등에 사용되고 있으며, 잘 구성된 온톨로지를 개발하는 것은 시간적, 물질적으로 많은 자원이 소모되는데, 자연어 문장의 서술어를 찾아 온톨로지 서술어로 자동 변환하는 방법도 제안되었다.[10]

3. 연구 방법

엔터테인먼트 분야에 국한된 자연어 검색시스템을 구현하기 위하여 우리는 몇 가지 단계의 처리로



[A typical natural language search scheme] [The proposed natural language search scheme]

Fig. 2. Comparison of Search Methods by way of Natural Language Input.

구성한다. 첫 번째로 정기적인 웹크롤링에 의해 엔터테인먼트 데이터만 가져와서 데이터베이스에 보관한다. 데이터베이스는 엔터테인먼트 데이터가 가장 잘 검색할 수 있도록 ERD를 구성한다. 두 번째로 엔터테인먼트용 질의어의 유형을 분석하여 패턴을 준비한다. 세 번째로 자연어 질의문장 입력에 대하여 격조사 인식위주의 간이 형태소 분석을 실시하고 가장 적합한 패턴을 매칭시킨다. 마지막으로 패턴에 대응하는 SQL문을 생성하여 데이터베이스에서 검색을 실시하고 결과를 돌려주도록 구성하였다.

정통적 방식의 자연어입력에 의한 검색은 자연어처리의 분석단계를 모두 거쳐야 한다. 이 때 대용량 사전과 한국어 문법이 필요하며, 분석 결과로 검색용 질의어를 생성하여 포털사이트를 검색하게 된다. 본 연구는 자연어입력에 대해 술어위주의 조사우선 간이형태소분석을 행하므로 사전이나 한국어 문법이 필요하지 않다. 분석결과를 준비된 패턴과 매칭하여 DB질의어를 생성하고 엔터테인먼트용DB로부터 검색결과를 도출해내는 것이 특징이다. Fig. 2에서 두 방식을 비교하였다.

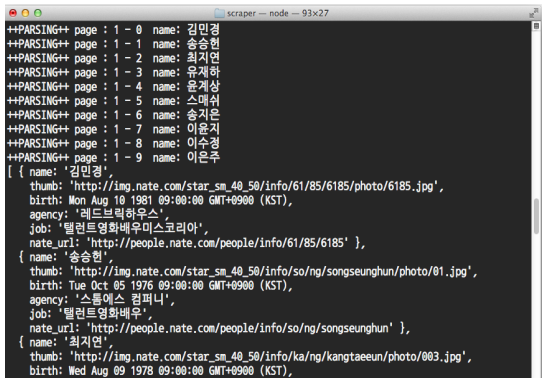


Fig. 3. An Example of Results of HTML Analysis.

3.1 웹크롤링에 의한 엔터테인먼트 데이터 수집

정해진 웹 문서를 DOM(Document Object Model) 또는 정규표현식을 통하여 분석하고, 분석된 데이터는 미리 정해진 키 테이블(Key table)을 참고한 맵핑(Mapping) 과정을 거쳐서 임시 저장소(Repository)에 저장한다. Fig. 3은 웹 문서 분석 결과를 보여준다.

키 테이블과 맵핑과정을 거치는 이유는, 동일한 인물에 대해 Fig. 4와 같이 생일 데이터를 "생몰" 로

Fig. 4. A Case of Inputting Different Keywords for an Identical Item.

Table 1. An Example of Key Table related to Personal Information

Key	Meaning	Format
name	'이름'	Text (String)
real_name	'본명', '실명', '진짜 이름'	Text (String)
birth	'생일', '출생일', '출생', '생몰'	YYYY '년' MM '월' DD '일' (Date)
birthplace	'고향', '출생지', '태어난 곳'	Text (String)
job	'직장', '직업'	Text, Text, Text (Array)
team	'소속그룹', '소속팀', '팀', '그룹'	Text, Text, Text (Array)
agency	'소속사', '소속', '회사'	Text, Text, Text (Array)
review	'데뷔', '첫 출연'	Date, Text (Array)
.....		

표기하고 있고, 다른 사이트에서는 "출생"으로 서로 다르게 표기하고 있지만 같은 의미를 가지고 있다. 이처럼 표기가 달라도 같은 데이터로 해석하기 위해 Table 1과 같은 키 테이블을 사용하여 맵핑한다. 또한, 질의문장의 표층에 나타나는 단어(랙심)를 인식하기 위하여 이들 Meaning 단어들을 형태소해석용 단어사전에 보관한다.

임시 저장소에는 데이터 갱신의 주기를 맞추기 위해 수집된 날자와 수집된 대상의 원본 URL(Uniform Resource Locator)을 가지고 있다. Fig. 5는 수집된 날짜의 reg_date 원본 URL을 nate_url로 정의하고 있다.

각각의 웹 사이트마다 수집되는 정보가 다르므로 사이트 마다 하나의 테이블이 필요하다. 이렇게 구성된 임시 저장소로 서비스를 구현하면 데이터의 중복성 문제와 데이터간의 관계형성이 어려워 관계형 데이터베이스로서의 이용 가치가 떨어지게 된다. 따라서 임시 저장소에 수집된 데이터를 재가공하여 주저장소의 데이터베이스로 만들어야 한다. Fig. 6은 엔터테인먼트관련 검색서비스를 제공할 데이터베이스의 설계도를 나타낸다. person 테이블을 중심으로 singer와 actor가 관계되어 있다. singer는 artist, song과, actor는 appearance, movie와 관계를 가진다. 보통은 1대1의 관계이지만 그룹으로 활동하는 경우는 artist에 그룹명을, singer에 복수의 그룹 멤버들을 보관한다.

3.2 단순한 질의문장 분석

엔터테인먼트 분야 중 음악관련 자연어 질의 312문장을 대상으로 유형을 파악하였으며, 아래의 Table

id	name	reg_date	thumb	birth	job	agency	nate_url
1	김세아	2012-11-04 02:05:51	http://img.nat...	1974-05-18 00:00:00	탤런트	휴메인엔터테인먼트	http://people.nate.com/people/info/ki/ms/kimseah
2	포지션	2012-11-04 02:05:51	http://img.nat...	1974-04-30 00:00:00	가수	스타폭스 엔터테인먼트	http://people.nate.com/people/info/35/66/3566
3	나나	2012-11-04 02:05:51	http://img.nat...	1991-09-14 00:00:00	가수	애프터스쿨오렌지 카...	http://people.nate.com/people/info/na/na/nana_af
4	김성수	2012-11-04 02:05:51	http://img.nat...	1968-10-03 00:00:00	가수	쿨JD 엔터테인먼트	http://people.nate.com/people/info/ki/ms/kimsungs
5	알리	2012-11-04 02:05:51	http://img.nat...	1984-11-20 00:00:00	가수	트로피엔터테인먼트예...	http://people.nate.com/people/info/al/li/alli
6	이기찬	2012-11-04 02:05:51	http://img.nat...	1979-01-10 00:00:00	가수	강동대학교호기심 스...	http://people.nate.com/people/info/le/ek/leekichar
7	윤진영	2012-11-04 02:05:51	http://img.nat...	1982-11-30 00:00:00	개그맨	QUAN 엔터테인먼트	http://people.nate.com/people/info/yu/nj/yunjin
8	박준형	2012-11-04 02:05:51	http://img.nat...	1969-07-20 00:00:00	가수영화배우	god	http://people.nate.com/people/info/pa/rk/parkjunhy
9	장미화	2012-11-04 02:05:51	http://img.nat...	1970-07-15 00:00:00	개그맨		http://people.nate.com/people/info/47/94/4794
10	박정수	2012-11-04 02:05:52	http://img.nat...	1953-06-01 00:00:00	탤런트영화배우	네오엔터테인먼트	http://people.nate.com/people/info/12/71/1271
11	송재희	2012-11-04 02:05:52	http://img.nat...	1979-12-11 00:00:00	탤런트영화배우	휴메인엔터테인먼트	http://people.nate.com/people/info/so/ng/songjaeh
12	션	2012-11-04 02:05:52	http://img.nat...	1972-10-10 00:00:00	가수	지누션YG엔터테인먼트	http://people.nate.com/people/info/se/an/sean
13	이준	2012-11-04 02:05:52	http://img.nat...	1988-02-07 00:00:00	가수영화배우	엠블랙제이튜컴프	http://people.nate.com/people/info/le/ej/leejunj
14	소유	2012-11-04 02:05:52	http://img.nat...	1992-02-13 00:00:00	가수	씨스타스타쉽엔터테인...	http://people.nate.com/people/info/si/st/sistar_soyu
15	김태훈	2012-11-04 02:05:52	http://img.nat...	1969-01-01 00:00:00	팝칼럼니스트DJ	TN엔터테인먼트	http://people.nate.com/people/info/ki/mt/kimteahu
16	이은정	2012-11-04 02:05:52	http://img.nat...	1985-02-07 00:00:00	탤런트모델	JYP엔터테인먼트	http://people.nate.com/people/info/67/86/6786
17	김호진	2012-11-04 02:05:52	http://img.nat...	1970-05-05 00:00:00	탤런트	태랑엔터테인먼트	http://people.nate.com/people/info/ki/mh/kimhojin

Fig. 5. An Example of Personal Data Collected in Temporary Storage.

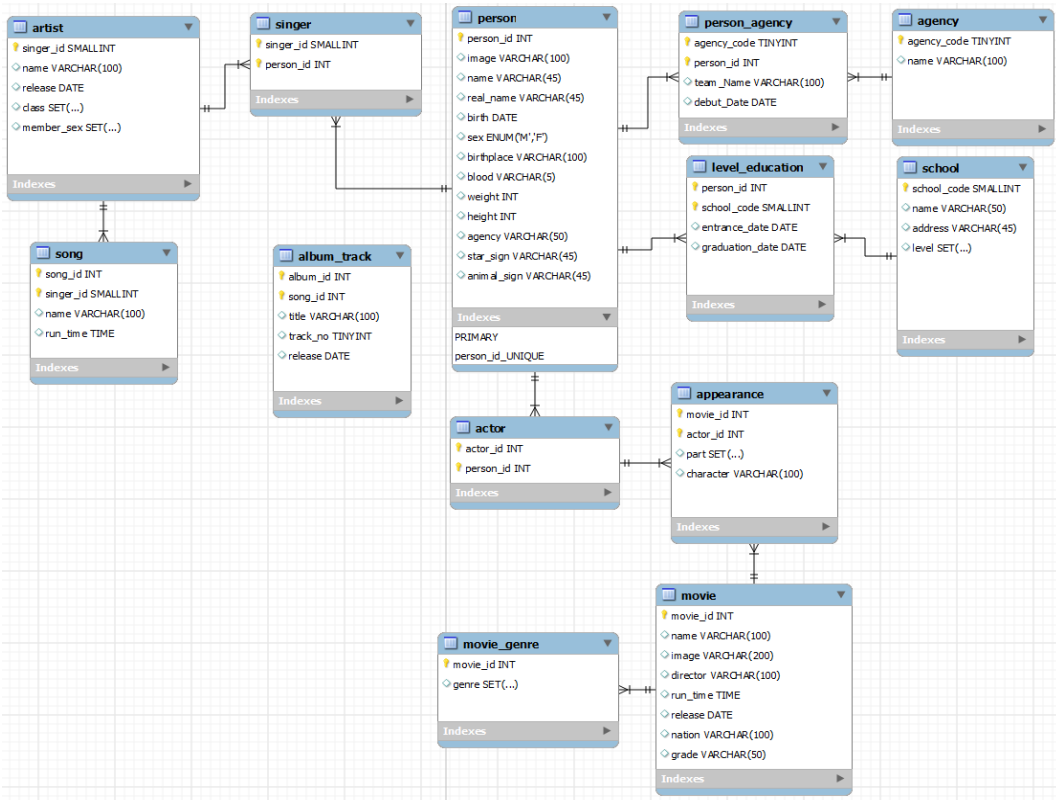


Fig. 6. Entertainment-Related Database ERD.

Table 2. Experimental Search Sentences

No	Search Sentences
1	소녀시대가 부른 노래는?
2	서른 즈음에를 부른 가수는 누구인가?
3	아이유와 김창완이 함께 부른 노래는?
4	아이유와 수지는 누가 나이가 많은가?
5	티아라와 멤버수가 같은 그룹은 어디인가?
6	수지와 같은 고향의 연예인은 누구인가?
7	한선화의 소속그룹은 어디죠?
	……
76	서른 즈음에는 누가 불렀나요?
77	부산이 고향인 가수는?
	……
168	김광석이 부른 '또 하루'로 시작하는 노래의 제목은?
	……
311	키가 180 이상이면서 고향이 서울인 가수는?
312	혈액형이 AB형인 가수는?

2는 검색 실험을 위한 실험 문장의 일부이다.

단순한 질의문장들은 술어가 없다. 즉, "수지의 고향은 어디인가", "김광석의 혈액형은 무엇인가", "아이유의 본명은 무엇이죠", "한선화의 소속그룹은 어디죠" 등은 단순한 질의문장이다. 이와 같은 형태는 주어진 단어(Given Word)와 찾고자 하는 key

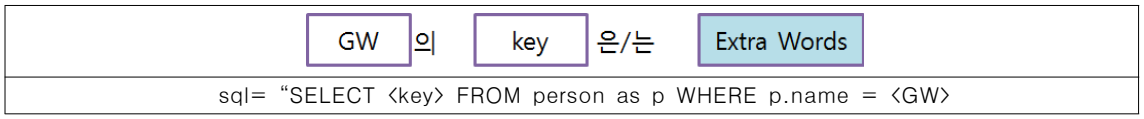


Fig. 7. Pattern of Simple Query Sentence Processing in the form of “<GW>의 <key>”.

(Meaning에서 key매칭)가 주어졌으므로 Fig. 7과 같은 검색패턴으로 검색이 가능하다.

술어가 없는 다른 질의문장으로 “부산이 고향인 가수는 누구인가”, “혈액형이 A형인 연예인은 누구죠”, “소속이 JYP인 가수는” 등의 형태가 있다. 우리는 Meaning으로 분류된 랙심(Lexeme)에서 key에 해당하는 단어를 찾을 수 있으므로 “고향이 부산인 가수”와 “부산이 고향인 가수”의 형태를 key(birth-place)가 앞쪽에 나타나는 패턴으로 준비하여 질의문장에 대해 적합한 답(Appropriate Answer)을 얻을 수 있다.

또한, 형용사와 부사가 나타난 경우로 “키가 160보다 작은 사람은?”, “수지보다 나이가 적은 가수는?”, “혈액형이 아이유와 같은 사람은 누구인가?”와 같은 종류의 문장인 경우는 Fig. 8의 “<GW> 인” 부분이 변형된 형태로 바뀌게 된다. SQL문은 질의의 결과를 다른 질의의 조건으로 활용하는 내포질의 형식이 된다.

3.3 질의 문장의 술어중심 분석

엔터테인먼트 분야에 국한될 경우 검색어의 술어는 “부르다, 노래하다, 춤추다, 공연하다, 연주하다, 발표하다, 출연하다, 등장하다, 나오다, 주연하다, 조연하다, 참여하다, 감독하다, 연출하다, 히트하다, 소속하다” 등으로 종류가 제한적이었다.

질의문장에 등장하는 술어가 제한적이므로 술어중심의 패턴을 통한 자연어처리가 가능하다. Table 2의 2번과 76번의 경우와 같이, “서른 즈음에를 부른 가수는 누구인가?”와 “서른 즈음에는 누가 불렀나요”는 그 뜻이 같은 질의문장이므로 정형화된 패턴으로 만들 수 있다면 동일한 SQL문을 생성하여 처리할 수 있다.

우선 우리는 음악과 관련된 질의에 초점을 맞추었다. 술어 “부르다”와 “노래하다”의 경우, 실제 질의문에서는 “부른, 불렀던, 부르는, 노래한, 노래하는” 등의 랙심들이 등장하였으며, 이들을 “부르다”의 변화형으로 술어사전에 모두 등록해두어 술어의 변화 형태에 관계없이 토큰 “부르다”로 매칭한다. 또한, “부르다”나 “노래하다”가 질의어 문장의 마지막에 등장할 경우는 “불렀나, 불렀습니까, 불렀나요, 불렀을까, 노래했나, 노래했나요, 노래했나여, 노래했습니까, 노래했을까요, 노래했을려나” 등과 같이 어미변화가

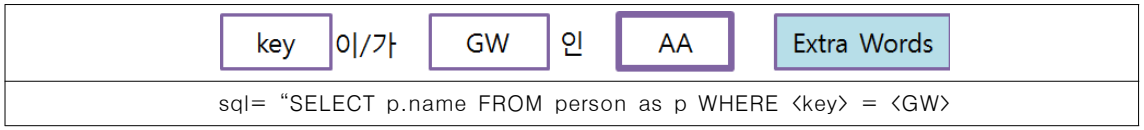


Fig. 8. Processing Pattern of Query Sentence Processing in the form of “<key>이/가 <GW> 인 <AA>”.

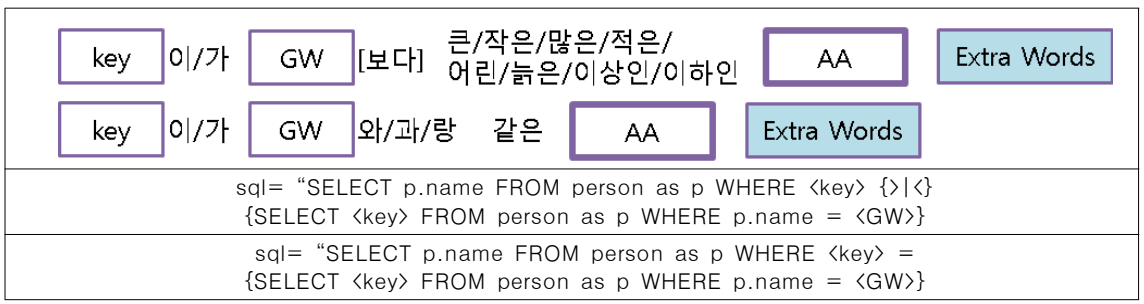


Fig. 9. Pattern of Query Sentence Processing including Adjective and Adverb.

Table 3. Recognition of the Token "부르다"

Type	Lexeme	Pattern	Token
Type 1 (middle of the sentence)	부른, 불렀던, 부르는, 노래한, 노래하는, 발표한	equal to Lexeme	부르다
Type 2 (end of the sentence)	불렀나, 불렀습니까, 불렀나요, 불렀을까, ...	불렀+ending	
	노래했나, 노래했나요, 노래했나여, 노래했습니까, 노래했을까요, 노래했을려나, ...	노래했+ending	

다양하게 일어나지만 "불렀"과 "노래했"이라는 공통적인 어간+과거형어미 형태의 랙심이 반드시 나타나는 경향을 보이기 때문에 변화형 어미를 무시한 "부르다"토큰의 생성이 가능하다 즉, 어미변화형 랙심들을 인지하더라도 우리는 Table 3과 같이 토큰 "부르다"를 생성할 수 있게 된다.

3.4 술어가 포함된 질의문의 패턴화

술어 중심의 질의문을 고려하면 다양한 형태의 질의문들 중에서 술어 "부르다"가 사용된 질의문들을 패턴으로 정형화할 수 있다. 결국 "부르다"가 가운데 포함되는 문장은 Fig. 10과 같은 격(case)형식으로 나타낼 수 있다.

"부르다"라는 술어가 가운데 표현된 Fig. 10의 정형화 문장은 앞쪽에 등장한 주격(가수)과 목적격(노래)이 주어졌을 경우, 뒤 쪽의 목적격이나 주격을 찾기 위한 질의문장의 형태이다. 뒤쪽에 배치되어 조사 "은/는" 다음에 나타난 "무엇"이나 "누구"와 같은 의문사는 질의문의 의미와 관계가 없으므로 버린다.

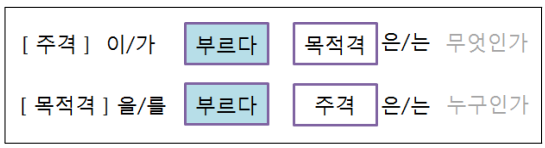


Fig. 10. Case Form of Query Sentence with the predicate "부르다" included in the middle of the Sentence (Type 1).

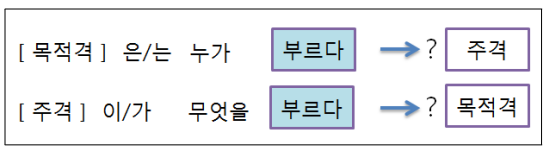


Fig. 11. Case Form of Query Sentence with the predicate "부르다" included at the end of the Sentence (Type 2).

Fig. 11은 "부르다"가 문장의 마지막에 나타난 경우로, 역시 목적격과 주격이 주어졌을 경우 주격(가수)이나 목적격(노래)을 찾는 행태가 된다.

Fig. 10과 Fig. 11에서 나타난 질의문들의 4가지 유형을 고려하여, "부르다" 토큰이 포함된 모든 질의문은 격조사를 참고하면 Fig. 12와 같은 패턴으로 정형화할 수 있다.

결국 토큰 "부르다"는 가수와 노래제목이 궁금한 경우에 사용하는 술어이며, 주격을 찾는 경우와 목적격을 찾는 경우가 있을 수 있다. 어느 쪽이던지 주어진 단어가 있고 찾고자 하는 정보가 있다. 우리는 자연어 질의에서 주어진 단어(GW)를 이용하여 거기에 가장 적합한 답변(AA)을 찾아서 돌려주어야 한다.

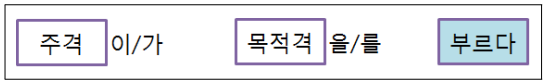


Fig. 12. Standardized Token of "부르다" by Case Particles.

3.4 격조사를 이용한 간이 형태소해석

3.4.1 주격조사에 의한 주격 질의문 처리

"서른 즈음에는 누가 불렀나요", "누가 서른 즈음에를 불렀더라", "서른 즈음에를 부른 사람은 누구인가" ,"서른 즈음에를 불렀던 가수가 누구지" 등의 문장은 격조사 "이/가","을/를", 보조사 "은/는"의 분석만으로 정형화된 패턴을 만들어낼 수 있다. 한국어는 어절별로 띄어쓰기를 하며, 특히 격조사는 어절의 마지막에 위치하는 특징이 있으므로 띄어쓰기가 된 어절의 마지막 부분에 격조사가 위치하는지를 확인하여, 그 앞쪽 단어의 격을 찾아낼 수 있게 된다.

서른 즈음에/ 는 / 누 / 가 / 불렀나요
.....(예 1)

누/ 가 / 서른 즈음에 / 를 / 불렀더라
.....(예 2)
서른 즈음에 / 를 / 부른 / 사람 / 은 / 누구인가
.....(예 3)
서른 즈음에 / 를 / 불렀던 / 가수 / 가 / 누구지
.....(예 4)

주격을 질의할 경우는 "누, 누구, 가수, 싱어, 사람"으로 표현되며, "이/가" 앞에 올 경우와 "은/는" 앞에 올 경우가 있다. 다만 예3의 마지막 어절 "누구인가"는 어절 마지막에 "가"가 위치했지만 격조사가 아닌 의문사로서의 역할을 담당하므로 문장에서 마지막 어절의 격조사 형태는 격조사로 인지하지 않는다. 표층에 나타난 "불렀xx"와 "부른, 불렀던"은 "부르다"로 변환하여 토큰이 생성되며, 격조사에 의해 정형화 처리를 하게 되면 위의 예제들은 모두 Fig. 13의 형태로 바뀌게 된다.

이제 "부르다"술어의 주어진 단어 GW는 "서른 즈음에"가 되며, 찾고자 하는 정보 AA는 가수 이름임을 알 수 있게 된다.

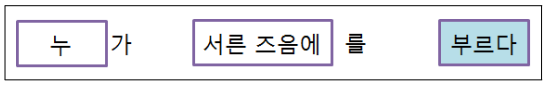


Fig. 13. Standardization of "부르다" Token in the form of Subjective Case Query.

3.4.2 목적격 질의문 처리

"부르다" 술어의 목적격에 해당하는 질의문을 분석한 결과, 아래와 같은 자연어 형태의 질의가 대표적이었다.

김광석 / 이 / 부른 / 노래 / 는
.....(예 5)
1995년에 / 김광석 / 이 / 부른 / 곡명 / 을 / 알려줘
.....(예 6)
김광석 / 이 / 불렀던 / 노래 / 중에서 / 또 / 하루로 / 시작하 / 는 / 노래 / 는 / 뭐지 .....(예 7)

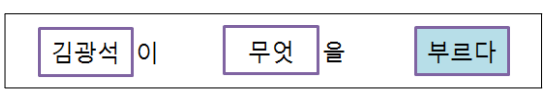


Fig. 14. Standardization of "부르다" Token in the form of Objective Case Query.

목적격 질의문의 특징은 주격 조사 "이/가" 앞에 가수명이 나타나며, 목격적조사 "을/를"이나 보조사 "은/는" 앞에 주로 "노래, 곡명, 제목, 노래명, 곡목, 노래제목"등의 단어가 나타나는 형태를 띄고 있다.

예 6과 같이 시기가 나타나는 경우, "년, 월, 년도, 연도"라는 특정 단어가 나타나며 부사구 형태를 띄게 되는데, 노래나 영화, 연극, 드라마 모두 데이터베이스에 일자가 함께 기록되어 있으므로 이를 검색에 이용하도록 한다.

3.4.3 주격이 복수인 목적격 질의문의 처리

술어 "부르다"에서 "아이유와 김창완이 함께 부른 노래는", "아이유랑 김창완이 함께 부른 노래제목이 뭐지"와 같이 주격이 복수인 경우에 목적격을 질의하는 형태이다. 접속조사 "와, 과,랑"이 "함께, 같이"라는 부사와 나타나는 경우가 일반적이었다.

주격에 해당하는 "아이유"와 "김창완"을 추출하면서 "함께, 같이"로 표현된 부사는 버려도 상관없다. 결국 정형화하면 Fig. 15와 같이 표현된다.

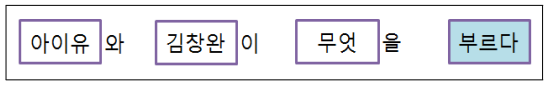


Fig. 15. Standardization of "부르다" Token in the form of Objective Case Query in the case of Subjective Case Being Plural.

3.5 패턴의 SQL문 생성

술어 "부르다"가 나타나는 자연어 문장을 처리하기 위하여 3가지 정형화 형태를 제시하였다. 정형화 형태를 패턴화하여 SQL문으로 생성한다면, 자연어 질의에 대해 데이터베이스를 검색하여 결과를 보여주는 것이 가능하다.

3.5.1 술어 "부르다"의 주격 질의

노래 제목으로 가수를 찾는 질의에 해당된다. 앞장에서 제시한 ERD의 song 테이블에서 노래 제목을 검색하여 singer테이블의 name필드를 출력한다.

"서른 즈음에를 부른 가수가 누구죠, 누가 서른 즈음에를 불렀더라, 서른 즈음에는 누가 불렀나요"등의 질의문은 모두 아래의 정형화 패턴으로 바뀌게 되었고, 이제 노래제목으로 가수 이름을 찾는 SQL만 대응시키면 된다.

3.5.2 술어 “부르다”의 목적격 질의

가수 이름으로 노래 제목을 찾는 질의에 해당된다. 즉, “김광석이 부른 노래는, 김광석이 불렀던 노래는 무엇이죠, 김광석은 어떤 노래를 불렀나요” 등과 같은 자연어 질의 문장이며, 앞 장에서 제시한 ERD의 singer테이블에서 가수 이름을 찾아서 해당 ID로 song 테이블에서 name필드를 출력한다.

3.5.3 술어 “부르다”의 기타형태 질의

가수 이름이 복수로 주어지는 경우의 질의이다. “아이유와 김창완이 같이 부른 노래제목이 뭐더라, 인순이와 조피디가 함께 불렀던 노래가 뭐죠” 등의 질의어 형태이며, AA=노래제목, GW=가수이름(“아이유, 김창완”), 혹은 GW=가수이름(“인순이, 조피디”)이 되어 토큰 “부르다”의 확장된 SQL문을 생성할 수 있다. 개인 아티스트와 그룹 아티스트를 관리하는 singer 테이블을 경유하므로 SQL문이 다소 복잡하게 보이지만, 우리는 “김창완”과 “아이유”를 아래의 SQL에 대입하도록 미리 형태를 만들어두어 처리하면 된다.

만약 3명의 이름으로 질의할 경우는 위의 SQL문 생성에 한사람 부분을 더 마련하여 질의어를 생성할 수 있도록 준비해야 한다. 일반화할 수도 있지만 3명 이상의 이름을 입력하면서 노래제목을 검색하는 경우는 거의 없으므로 3명까지만 처리하도록 질의어 패턴을 준비하였다.

3.6 검색결과의 고찰

노래와 가수 관련 질의문장과 토큰 “부르다” 관련

누 가 서른 즈음에 를 부르다
Token= “부르다”, GW=노래제목(“서른 즈음에”), AA=가수이름
SELECT a.name as artist_name FROM song as so, artist as a WHERE so.name= ‘서른즈음에’ AND a.singer_id=so.singer_id;
sql= “SELECT a.name as artist_name FROM song as so, artist as a WHERE so.name= ‘” + gw + “’ AND a.singer_id=so.singer_id;” ;

Fig. 16. Pattern of SQL Query for GW=노래제목(“서른 즈음에”) of “부르다”.

김광석 이 무엇 을 부르다
Token= “부르다”, GW=가수이름(“김광석”), AA=노래제목
SELECT so.name FROM song as so, artist as a WHERE a.name= ‘김광석’ and a.singer_id=so.singer_id;
sql= “SELECT so.name FROM song as so, artist as a WHERE a.name= ‘” + gw + “’ and a.singer_id=so.singer_id;

Fig. 17. Pattern of SQL Query for GW=가수이름(“김광석”) of “부르다”.

Token= “부르다”, GW=가수이름(“아이유,김창완”), AA=노래제목
SELECT so.name as song_name, tmp.name as artist_name FROM song so
JOIN(SELECT a.singer_id, a.name FROM singer si LEFT OUTER JOIN person p ON p.person_id=
si.person_id JOIN artist a on si.singer_id=a.singer_id
WHERE p.name LIKE ‘%김창완%’ or p.name LIKE ‘%아이유%’
GROUP BY si.singer_id HAVING count(si.singer_id)>=2) tmp
ON so.singer_id=tmp.singer_id;”

Fig. 18. Pattern of SQL Query for GW=가수이름(“아이유, 김창완”) of “부르다”.

```
Token= “부르다” , GW=가수이름( “S1, S2, S3” ), AA=노래제목
sql= “SELECT so.name as song_name, tmp.name as artist_name FROM song so JOIN(SELECT
a.singer_id, a.name FROM singer si LEFT OUTER JOIN person p ON p.person_id=
si.person_id JOIN artist a on si.singer_id=a.singer_id
WHERE p.name LIKE ‘%” + S1 + “%’ or p.name LIKE ‘%” + S2 +
“%’ or p.name LIKE ‘%” + S3 + “%’ GROUP BY si.singer_id HAVING count
(si.singer_id)>=3) tmp
ON so.singer_id=tmp.singer_id;”
```

Fig. 19. Pattern of SQL Query for GW=가수이름(“S1, S2, S3”) of “부르다”.

질의문장들을 자연어 입력하여 테스트한 결과, 312문장 중에서 298문장의 처리가 원만하게 이루어졌다. 질의문은 “가수와 노래에 대하여 궁금한 사항을 검색용 질의문으로 자유롭게 작성하라”는 지침으로 15명이 약 20문장씩을 작성하였다. 처리결과는 각 질의와 결과에 대하여 5명이 성공과 실패로 분류하였으며, 3명 이상이 성공이라 표시한 결과를 적절한 응답으로 분류하였다. Table 4는 실험결과를 나타낸다.

“김광석이 부른 ‘또 하루’로 시작하는 노래의 제목은?” 같은 경우는 현재의 구조에서 처리할 수 없었으며, 검색결과로 김광석의 노래들이 모두 출력된다. 개선 시스템으로 노래 첫 소절을 song 테이블에 하나의 필드로 저장한다면 검색이 가능하겠지만, 가사 중에 “내가 떠나보낸 것도 아닌데”와 같은 특정 내용이 들어간 노래를 찾는 경우까지 생각한다면 모든 가사를 저장해야 하는 부담이 있다. 또한, “걸스데이 민아와 손흥민이 언제부터 사귀기 시작했는가” 등과 같은 최근 이슈들도 현재 구조로는 처리할 수 없다. 이를 위해서는 이슈 처리를 위한 특별한 구조의 테이블이 별도로 준비되어야 한다.

Table 4. Experimental Results of 321 Query Sentences

Number of people (appropriate response)	A typical natural language search scheme (cumulative %)	The proposed natural language search scheme (cumulative %)
5 persons	195 (62%)	220 (71%)
4 persons	67 (84%)	56 (88%)
3 persons	25 (92%)	22 (96%)
2 persons	16 (97%)	14 (100%)
1 person	9 (100%)	0 (100%)
0 person	0 (100%)	0 (100%)
Total	312	312

정상적인 포털검색, 자연어 검색, 술어 기반 자연어 처리와 본 논문에서 제안하는 엔터테인먼트 분야에 국한된 자연어 검색방법은 다음과 같은 차이가 있다.

엔터테인먼트 분야에 국한하여 자연어 검색을 시도한 Table 5의 비교결과를 보면, 시스템의 복잡도는 감소하면서 검색결과의 적절한 응답률은 올라갔다는 것을 알 수 있다. 또, 본 검색 시스템은 질의 결과로 다시 질의하여 결과를 도출하는 2차 질의가 한 번의 처리로 가능하였다. 예를 들면 예전의 검색 시스템들은 “수지와 같은 고향의 연예인은”과 같은 질의의 경우, “수지의 고향은?”을 질의하여 그 결과 “광주”를 가지고 다시 “고향이 광주인 연예인은”이라 검색해야 한다. 위의 질문에 대하여 Fig. 8의 패턴을 활용하면 한 번의 질의로 적합한 답을 얻을 수 있게 된다.

4. 결 론

최근 한류가 세계적으로 큰 관심을 끌고 있어서, 포털사이트 검색의 상당부분이 엔터테인먼트와 관련되어 있다. 하지만, 단순한 키워드 검색이 대부분이며 자연어 질의에 대한 처리는 많이 미흡하다. 우리는 광범위한 엔터테인먼트 데이터에 대해 웹클로링으로 3만2천여 건의 관련데이트를 수집하여 데이터베이스화하고 자연어 질의로 원하는 결과를 얻을 수 있도록 엔터테인먼트전용 검색시스템을 구축하였다. 그리고 엔터테인먼트 분야에 국한되지만 최소한의 자연어 분석으로 최대의 효과를 얻을 수 있는 간이 자연어 질의응답 시스템을 구축하였으며, 영화, 드라마보다 가수와 노래에 대해 술어중심 자연어 검

Table 5. Comparison of Characteristic Points Consequential on Search Methods

Items of Comparison	Search in General Portals	General Natural Language Search	Predicate-based Natural Language Processing	Search method proposed in the paper
Search Method	Keyword Input	Natural Language Input	Natural Language Input	Natural Language Input for Entertainment Data
Search Target	Web documents, blogs, etc.	Web documents, blogs, etc.	Web documents, blogs, etc.	Use of specialized entertainment DB
Data Collection and Structure	Indexing of web documents, blogs, etc. by periodic web-crawling	Index structure or general portals and mass storage dictionary for natural language processing	Index structure or general portals and mass storage dictionary for natural language processing	DB update after periodic web-crawling of entertainment data
Processing Complexity	Indexing of massive data	Requires authentic natural language processing utilizing a dictionary by POS(part-of-speech)	Reduction of complexity in terms of predicate-based natural language search	Predicate-oriented Small-scale natural language processing of entertainment data
Search Result	Chooses the needed results among many search results	Chooses the needed results among many search results	Chooses the needed results among many search results	Draws the accurate results because it is specialized for searching entertainment data
Availability of Secondary Query	Twice	Twice	Twice	Only Once
Response Rate to Query	92%(287/312)	-	-	96%(298/312)

색이 가능하도록 검색시스템을 구축하였다. 그 결과, 가수와 노래에 대한 질의는 96%이상 원하는 결과를 찾을 수 있었다.

본 연구는 자연어 질의를 대상으로 하는 일반적인 형태소분석과 구문분석 절차에 비해 사전이나 한국어 문법이 필요 없는 격조사 일부를 이용하는 간이 형태소 분석을 수행하므로 자연어 입력문을 분석하는 과정이 크게 간소화 되었다. 형태소분석 결과를 활용하여 술어 중심의 정형화된 패턴 추출과 SQL 문 매칭만으로 검색 처리가 가능하며, 그러한데도 비슷한 성능을 낼 수 있다는 데에 큰 의미가 있다. 다만, 엔터테인먼트와 관련된 많은 자연어 입력문장들을 대상으로 검색 패턴을 만들고, 유사한 패턴들을 통합해 나가는 과정들이 추가적으로 필요하다. 앞으로 영화와 드라마 등에서도 활용할 수 있도록 술어 중심의 패턴을 늘려서 전반적 엔터테인먼트 분야에서 통용될 수 있는 자연어 질의 검색 시스템으로 확장해 나갈 예정이다.

REFERENCES

[1] Tour Information Retrieval Application, Desti, Smart Tour Guide combine Natural Language Retrieval and Artificial Intelligence, http://techneedle.com/archives/5979, 2012.

[2] S.E. Shin and Y.H. Seo, “Predicate-based Question Analysis for Korean Question-answering System,” *Journal of the Research Institute for Computer and Information Communication*, Vol. 13, No. 1, pp. 99-104, 2005.

[3] S.H. Chung and D.S. Hwang, “A Korean Morphological Analysis System for Korean Spoken Sentences,” *Proceeding of Conference of the Korean Institute Scientists and Engineers*, Vol. 25, No. 1B, pp. 414-416, 1998.

[4] S.E. Shin and Y.H. Seo, “Question Analysis based Syntactic Information in Korean Question Answering System,” *Proceeding of Conference of the Korean Institute Scientists*

and Engineers, Vol. 31, No. 1B, pp. 931-933, 2004.

[5] W. Lee, J.J. Song, and M.M. Kang, "Effective Patent Retrieval Using Query Relationship," *Proceeding of Conference of the Korean Institute of Industrial Engineers*, Vol. 2012, No. 5, pp. 388-395, 2012.

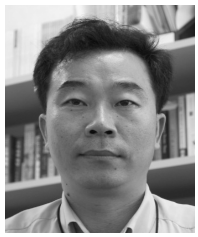
[6] S.E. Shin and Y.H. Seo, "Deep Analysis of Question for Question Answering System," *Journal of the Korea Contents Society*, Vol. 6, No. 3, pp. 12-19, 2006.

[7] H.S. Kim, Y.H. An, and J.Y. Seo, "A Question Type Classifier based on a Support Vector Machine for a Korean Question-Answering System," *Journal of the Korean Institute Scientists and Engineers*, Vol. 30, No. 5, pp. 466-475, 2003.

[8] G.C. Kim, "Korean Language Query Analysis With an Interrogative Pronoun For Information Retrieval," *Journal of The Korean Institute of Communication Sciences*, Vol. 33, No. 2D, pp. 48-54, 2008.

[9] J.S. Chae and S.H. Lee, "Design and Implementation of Korean Query System using Frame-based Query Decomposition Method," *Journal of the Korean Institute Scientists and Engineers*, Vol. 4, No. 3, pp. 452-461, 1997.

[10] Y.K. Min and B.J. Lee, "Automatic Ontology Generation from Natural Language Sentences Using Predicate Ontology," *Journal of Korea Multimedia Society*, Vol. 13, No. 9, pp. 1263-1271, 2010.



김 정 인

1991년 4월~1993년 3월, 게이오대학 계산기과학전공 공학석사

1993년 4월~1996년 3월, 게이오대학 계산기과학전공 공학박사

1996년 5월~1998년 2월, 포항공과대학교 정보통신연구소 연구원, 기계번역시스템 설계

1998년 3월~현재, 동명대학교 컴퓨터공학과 교수

관심분야 : 기계번역, 기계학습, 시멘틱웹, 웹2.0