온톨로지 지식기반 질의 의미 해석 검색

김난주, 정훈, 표혜진, 최의인
한남대학교 컴퓨터공학과
e-mail : 91knj@naver.com

Ontology Knowledge-based query semantic analysis search

Nanju Kim, Hoon Jeong, Hyejin Pyo, Euiin Choi
Dept. of Computer Engineering Hannam University Daejeon, Republic of Korea

요 약

시맨틱 검색은 논리적으로 표현된 지식 베이스를 사용하여 현재의 키워드 기반 검색보다 더 정확한 결과를 제공할 수 있다. 그러나 일반 사용자는 지식 기반의 복잡하고 정형화된 질의어와 스키마를 잘 알지 못한다. 그래서 검색 시스템은 사용자 키워드의 의미를 해석할 수 있어야 한다. 본 논문에서는 멀티미디어 콘텐츠의 시맨틱 검색을 위한 사용자 질의 의미 해석 시스템을 설명한다. 제안한 시스템은 도메인 온톨로지 기반으로 구축된 지식 베이스의 정형화된 구조에 의미 해석 과정이 통합된 온톨로지 지식 베이스 기반 검색 시스템이다.

1. 서론

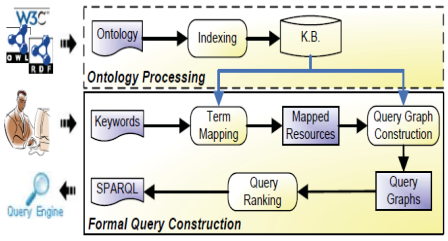
시맨틱 검색(Semantic Search)은 검색 결과의 정확도를 향상시키기 위해 기존의 키워드 기반 정보 검색(Information Retrieval) 알고리즘 방식을 탈피하여 능동적으로 사용자의 의도를 파악하고, 기존 정보를 가공 분석하여 정교한 검색 결과를 도출하는 일련의 활동 및 방법론을 통칭한다[1]. 최근에는 국내의 네이버, 다음, 네이트와 같은 포털 사이트의 검색 엔진들도 시맨틱 검색 기술을 도입하고 상용화하기 위한 노력을 기울이고 있다[2].

이처럼, 시맨틱 검색이 정확도 높은 검색을 제공하기 위해서는 첫 번째로 사용자로부터 입력된 부정확한 질의어의 의미를 정확하게 해석하기 위한 방법이 필요하다. 키워드 기반 검색에 익숙한 사용자들은 지식 베이스 기반의 정형화된 질의어와 스키마에 대한 이해가 없기 때문에 기존 키워드 기반의 검색처럼 몇 개의 키워드만으로 검색을 수행하게 된다[6, 8]. 두 번째로 잘 구축된 풍부한 지식 베이스가 필요하다. 하지만, 지식 베이스의 구축은 도메인이 한정된다 하더라도 쉽지 않은 일이다. 특히, 실시간 이슈성 키워드와 같이 시간에 따라 변화하는 키워드들을 지식 베이스에 반영하는 것은 더욱 어려운 일이다[3].

따라서 본 논문에서는 정확도 높은 시맨틱 검색을 제공하기 위하여 사용자 검색문의 정확한 의미를 해석하기 위한 온톨로지 기반 사용자 질의 의미 해석 기법을 제안한다. 제안한 질의 의미 해석 기법을 통해 부족한 사용자 검색문으로부터 정확한 검색 의도를 해석할 수 있다.

2. 관련 연구

SPARK 시스템은 사용자 검색 키워드를 지식 베이스 기반으로 의미를 해석하고, 해석된 결과로부터 시맨틱 검색을 위한 정형화된 질의어를 구성하여 검색 결과를 제공하는 시맨틱 검색 시스템이다[4, 5]. (그림 1)은 SPARK 프레임워크의 구조를 설명하고 있다. SPARK 시스템은 크게 3 가지 단계를 통해 사용자 검색 키워드의 의미를 해석하고, SPARQL[8]과 같은 정형화된 질의를 생성하여 검색 결과를 제공한다. SPARK 시스템은 키워드 기반 검색에 익숙한 사용자들이 정형화된 질의의 표현 방법을 전혀 모르는 상태에서 일반 키워드 검색과 동일한 검색 키워드를 제공하고 SPARK 시스템이 이를 해석하여 질의 조합을 구성하는 것이 특징이다.



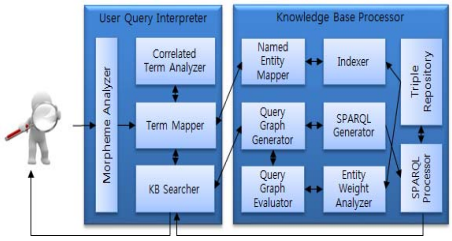
(그림 1) SPARK 프레임워크

- 개체 식별(Term Mapping): 사용자 키워드로부터 지식 베이스 내의 지식 개체를 식별
- 질의 그래프 생성(Query Graph Construction): 식별된 지식 개체를 바탕으로 시맨틱 네트워크를 탐색하여 질의 그래프를 구성
- 질의 랭킹(Query Ranking): 생성된 질의 그래프를

사용자 질의와 지식 베이스를 기반으로 평가

3. 사용자 질의 의미 해석 시스템

시맨틱 멀티미디어 콘텐츠 검색을 위한 시스템의 구조는 (그림 2)에서 보는 바와 같이 크게 사용자 검색문 의미 해석기와 지식 베이스 처리기로 구성된다. 지식 베이스 처리기는 사용자 검색 키워드를 지식 베이스 기반으로 개체 식별하고, 식별된 개체로부터 다수의 후보 질의 그래프를 구성한다. 구성된 다수의 질의 그래프를 평가하여 사용자 검색 의도에 가장 부합하는 질의 그래프를 찾고 질의 그래프를 SPARQL로 변환하여 검색 결과를 사용자에게 제공한다. 사용자 검색문 의미 해석기는 지식 베이스 처리기를 활용하여 사용자 검색문의 의미를 해석한다.

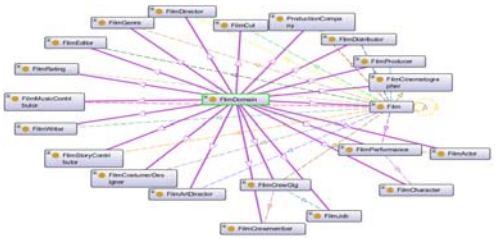


(그림 2) 시맨틱 멀티미디어 콘텐츠 검색 구조

3.1 온톨로지 기반 지식 베이스

시맨틱 검색에 있어 가장 전제되어야 할 것은 잘 구축된 지식 베이스이다. 제안 시스템에서는 지식 베이스 구축을 위해 프리베이스(freebase)를 참조하여 온톨로지 스키마를 설계하였다. 온톨로지 설계 참조 모델의 선정 기준은 스키마의 지식 표현 범위, 지식 표현 방법, 지식의 풍부성을 고려하였다. 프리베이스의 경우, OWL(Web Ontology Language)[7]을 이용하여 직접 표현 가능하고, 영화, 방송(예능, 드라마), 뮤직비디오 콘텐츠 검색을 위한 프로토타입 시스템의 지식 표현 범위를 충분히 만족하고 있다. 온톨로지 모델 설계를 위해 BBC 프로그램, IMDb 등의 온톨로지를 분석하였으나, 특정 도메인의 지식만을 표현하기 위해 설계되어 있으며, 상이한 표현 방법을 사용하고 있기 때문에 이들의 통합이 어렵다. 프리베이스에서는 타입(type)을 특정 토픽에 대한 is-a 관계로 나타내며, 타입의 종류는 열거형 타입(enumerated type), 상속형 타입(included type), 다중 필드 표현을 지원하는 CVT(Compound Value Type)으로 구성된다. 프리베이스에서 속성은 쌍으로 연결된 두 개의 타입의 관계를 나타내며, 각 속성의 아크는 항상 단방향성 아크를 의미한다. 아크의 네이밍룰은 아크가 속한 영역의 아이디, 아크가 시작되는 쪽의 클래스 아이디에 아크(속성)의 이름을 붙임으로써 아크 ID 의 유일성을 유지한다.

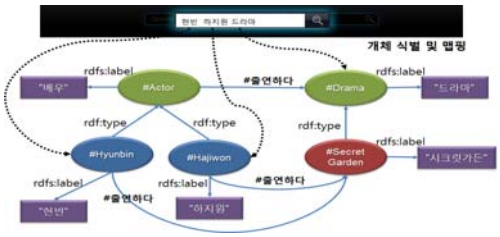
(그림 3)은 영화 도메인을 구성하는 각 클래스를 보여주고 있다.



(그림 3) 영화 도메인의 하위 클래스

3.2 개체 식별

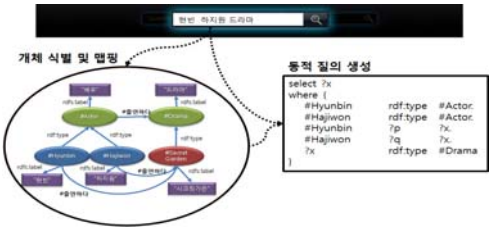
개체 식별 과정은 사용자로부터 입력받은 검색 키워드를 분석하여 지식 베이스 내에서 존재하는 지식 개체를 찾는 과정이다. (그림 4)는 개체 식별 및 매핑 과정을 설명하고 있다. 사용자가 “현빈 하지원 드라마”라는 검색 키워드를 입력하게 되면, 시스템은 형태소 분석을 통해 입력된 키워드를 분리하고 “현빈”, “하지원”, “드라마”의 각 키워드에 해당하는 지식 개체를 지식 베이스로부터 식별한다.



(그림 4) 개체 식별

3.3 질의 그래프 생성 및 평가

개체 식별이 완료되면, 지식 베이스로부터 식별된 개체를 모두 포함하는 부분 그래프를 찾는 질의 그래프 생성 과정을 수행한다. 부분 그래프를 찾는 과정에서 식별된 지식 개체를 포함하는 부분 그래프는 다양하게 존재 가능하다. 이러한 부분 그래프는 사용자가 입력한 검색 키워드로부터 해석 가능한 다양한 후보 해석 대안이 될 수 있으며, 이러한 후보 해석 대안을 지식 베이스 기반으로 평가하여 가장 적합한 해석 대안을 결정하고, 결정된 해석 대안으로부터 SPARQL 형태의 질의문을 생성한다. (그림 5)는 질의 그래프 생성을 나타낸다.



(그림 5) 질의 그래프 생성

4. 시스템 검증 및 실험

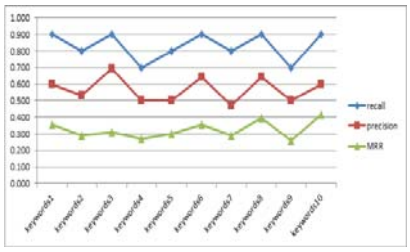
4.1 실험환경 구성

제안된 검색 시스템의 검증을 위해 10 개의 검색 예제를 정하였다. 그리고, 재현율(recall rate)과 정확율(precision rate) 평가를 위해 영화, 드라마, 예능 콘텐츠 기준으로 각 검색 예제 별로 10 개의 적합 콘텐츠 집합을 선별하고, 10 개의 부적합 콘텐츠 집합을 선별하였다. 또한 각 예제의 정답 집합으로부터 MRR(Mean Reciprocal Rank) 평가를 실험하였다. 예제별로 TF/IDF 기반의 키워드 검색 방법과 제안 시스템에서 검색을 수행하여 검색 성능을 평가하였다.

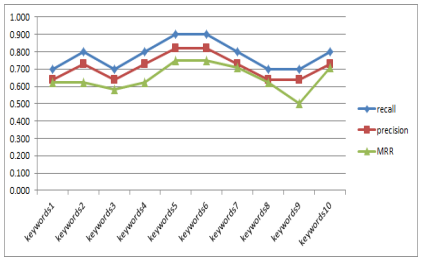
4.2 실험 결과

(그림 6)과 (그림 7)은 각 시스템에 대하여 검색 예제별로 평가한 재현율, 정확율, MRR 을 설명하고 있다. 실험결과에서 보는 바와 같이 TF/IDF 기반 검색은 다른 시스템에 비하여 재현율은 높으나 정확율이 급격히 감소하고 있음을 알 수 있다. 그리고 기법의 특성상 타 시스템에 비하여 재현율이 높다. 더욱이 TF/IDF 기반 시스템의 경우 MRR 이 재현율에 비해 현격히 낮은 성능을 보여주는 것을 알 수 있다. 검색 결과에는 포함되어 있지만, 검색 결과가 나타나는 순위가 낮음을 보여주는 것이다.

논문에서 제안하는 사용자 질의 확장 검색의 경우, 기존 시스템과의 객관적인 성능 평가 및 비교는 불가능하다. 하지만, 제안 시스템에서는 질의 확장기를 통하여 정확한 검색 결과를 제공한다.



(그림 6) 검색 키워드에 대한 실험 결과 (TF/IDF)



(그림 7) 검색 키워드에 대한 실험 결과 (제안 시스템)

5. 결론

본 논문에서는 온톨로지 지식 베이스를 기반으로 사용자 질의 의미 해석과 확장 기법을 제안하였다. 또한, 제안된 기법에 따라 프로토타입을 구현하고 실험을 통하여 TF/IDF 기반의 키워드 기반 검색과 타 시맨틱 검색 시스템보다 정확한 검색 결과를 제공하는 것을 확인 할 수 있었다. 제안된 시스템은 한정된 도메인에 대한 지식 베이스를 구축하고 프로토타입 형태로 검색 결과 실험을 수행하였으나, 검색 성능에 있어 타 시스템과 비교하여 우수한 성능을 제공하고 있다.

향후 연구 과제로 사용자 검색 패턴 분석을 통해 사용자 선호 기반 질의 그래프 평가 기법에 대한 연구를 진행 할 것이다.

감사의 글

본 연구는 교육부와 한국연구재단의 지역혁신인력양성사업으로 수행된 연구결과임

본 연구는 미래창조과학부 및 정보통신산업진흥원의 IT/SW 창의연구과정의 연구결과로 수행되었음 (NIPA-2013-1103)

참고문헌

[1] 김정민, 정현숙, “방송 온톨로지 구축 및 매칭 기반의 방송 프로그램 검색”, 한국정보기술학회 제 9 권 제 12 호, pp.161-171, 2011 년 12 월.

[2] 정휘웅, 김경선, 정한민, “시맨틱 검색 기술 동향”, 주간기술동향 통권 1431 호, 정보통신산업진흥원, 1432 호, pp. 14-27, 2010 년 2 월.

[3] 이동균, 권준희, “최근 사용자 관심사를 고려한 소셜 검색 알고리즘”, 한국정보기술학회 제 9 권 제 4 호, pp.187-194, 2011 년 4 월.

[4] Q. Zhou, C. Wang, M. Xiong, H. Wang and Y. Yu, “SPARK: Adapting Keyword Query to Semantic Search”, LNCS vol. 4825, 2007

[5] T. Tran, P. Cimiano, S. Rudolph and R. Studer, “Ontology-Based Interpretation of Keywords for Semantic Search”, LNCS vol.4825, 2007

[6] E. Mäkelä, “Survey of Semantic Search Research”, in Proceedings of the Seminar on Knowledge Management on the Semantic Web, 2005

[7] OWL: Web Ontology Language Reference (http://www.w3c.org/TR/owl-ref/), W3C, 2004

[8] Y. Lei, V. Uren and E. Motta, “SemSearch: A Search Engine for the Semantic Web”, LNCS(LNAI) vol.4248, 2006.