시맨틱 콘텐츠 검색을 위한 질의 확장 시스템

이무훈*, 최의인**

Query Expansion System for Semantic Contents Retrieval

Moo-Hun Lee*, Eui-In Choi**

요 약 최근 논리적으로 표현된 지식 베이스를 사용하는 키워드 기반 검색에서 보다 더 정확한 결과를 제공하기 위해 시맨틱 검색 방법에 대한 연구가 진행되고 있다. 대부분의 사용자는 정형화된 질의어와 스키마를 사용하는 것 보다 사용자 키워드의 의미를 해석해서 사용한다. 본 논문에서는 시맨틱 검색을 위한 사용자 질의 확장을 제안한다. 제안 시스템에서는 지식 베이스와 연관 검색어를 활용한 사용자 질의 확장 콤포넌트와 사용자 질의 해석 결과를 조정하기 위한 콤포넌트를 제공한다. 마지막으로 논문에서 제안한 사용자 질의 의미 해석 기법의 검증을 위해 프로토타입 시스템의 실험 결과를 설명한다.

주제어 : 시맨틱 검색, 온톨로지, 지식베이스, 검색어 확장 시스템

Abstract For semantic search methods to provide more accurate results than keyword-based search in a logical representation that uses a knowledge base are being studied. Than most of the user to use formal query language and schema used to interpret the meaning of a user keyword. In this paper, we propose to expand the user query for semantic search. In the proposed system, user query expansion component and a component to adjust the results to interpret user queries to take advantage of the knowledge base associated with a search term. Finally, a user query semantic interpretation, the proposed scheme to verify the experimental results of the prototype system is described.

Key Words : Semantic Search, Ontology, Knowledge Base, Query Expansion System

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 시맨틱 검색의 기술 성숙도가 높아짐에 따라 마이크로소프트 Bing, 퀀투라(Quintura), 볼프럼 알파(Wolfram Alpha) 등 국외 검색 서비스에서 시맨틱 검색 기술을 도입하고 있다. 뿐만 아니라, 국내의 네이버, 다음, 네이트와 같은 포털 사이트의 검색 엔진들도 시맨틱 검색 기술을 도입하고 상용화하기 위한 노력을 기울이고 있다[2, 3].

최근에는 스마트폰, 스마트TV 등과 같이 PC 형태의 디바이스에서 다양한 형태의 디바이스로 발전되어 가고 있다. 이러한 디바이스들은 인터넷에 항상 연결되어 있으며, 다양한 형태로 정보 검색을 시도하고 있다. 이러한 디바이스의 특징 중 하나는 휴대 편의성을 통해 언제 어디서나 검색이 가능한 반면에 입출력이 제한적이다. 종래 PC 환경에서의 정보 검색은 편리한 입출력 장치를 통해 다양한 키워드로 검색을 수행하고, 그 결과를 사용자가 직접 네비게이션하며 검색 결과를 획득하고 있다. 하지만, 스마트폰, 스마트TV와 같은 환경에서는 기존 PC 환경과 같은 입출력 장치를 제공하는데 한계가 있기 때문에 정보 검색에 있어서도 부족한 정보를 잘 해석하여

*한국전자통신연구원
**한남대학교 컴퓨터공학과 교수(교신저자)
논문접수: 2012년 10월 19일, 1차 수정을 거쳐, 심사완료: 2012년 11월 14일

사용자가 원하는 정확한 검색 결과를 제공하여야 한다. 즉, 재현율 위주의 기존 키워드 기반 정보 검색에서 정확도 위주의 시맨틱 검색 환경으로 변화하고 있다.

본 논문에서는 정확도 위주의 시맨틱 검색을 제공하기 위하여 이슈성 키워드와 같이 지식 베이스로 구축하기 어려운 메타데이터를 연관검색어 기반으로 확장하는 사용자 질의 확장 기법을 제안한다. 또한 질의 확장 기법을 통해 실시간성 이슈가 되는 키워드들을 지식 베이스에 업데이트 하지 않고도 사용자 검색문의 의미를 확장하여 검색의 정확도를 높일 수 있도록 하였다[7].

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대해 소개하고, 3장에서는 제안한 사용자 질의 의미 해석 시스템에 필요한 질의 확장 시스템을 설명한다. 4장에서는 제안 시스템의 성능 평가 및 분석에 대해 설명하고, 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

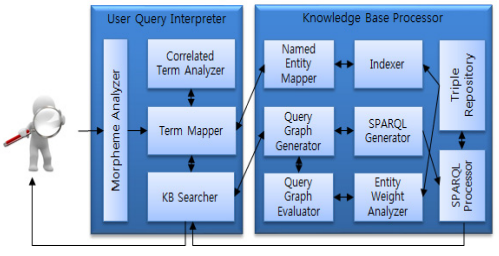
네이버의 연관검색어 서비스는 사용자가 많이 찾는 검색어를 분석하여 시스템에 의해 자동 노출되는 서비스이다[6, 8]. 그림 1과 같은 네이버 연관검색어의 기본 알고리즘은 사용자가 “a”라는 키워드로 검색을 수행한 후, “b”라는 검색 키워드로 검색한 회수를 측정하여 산출한다. 이러한 알고리즘에는 사용자가 직접 검색 키워드를 입력하고, 다시 두 번째 검색 키워드를 입력하게 되면 첫 번째 검색 키워드와 두 번째 검색 키워드 간의 연관관계가 있다고 가정한다. 또한 사용자가 최초 입력한 검색 키워드로 검색을 수행하고 네이버에서 연관검색어를 노출해주면 사용자는 제시된 연관검색어를 통해 다시 검색을 수행하는 경우, 최초 입력된 사용자 검색 키워드와 선택한 연관검색어 사이의 연관도가 증가하는 방식을 사용하고 있다. 간단한 알고리즘이지만, 네이버가 국내 검색 엔진 중에서 가장 높은 점유율을 차지하고 있다는 측면에서 많은 사용자로부터 입력되는 1차, 2차 검색 키워드의 패턴을 분석한 연관검색어 패턴 데이터는 검색에 있어 상당히 가치가 있는 데이터라 할 수 있다.

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] 네이버 연관검색어

3. 사용자 질의 확장 시스템

시맨틱 멀티미디어 콘텐츠 검색을 위한 시스템의 구조는 그림 2에서 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[9, 10].



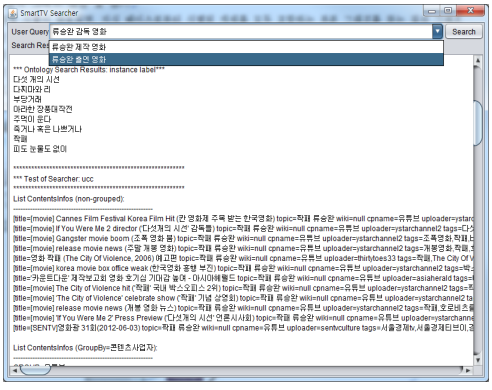
[그림 2] 시맨틱 멀티미디어 콘텐츠 검색

앞 장에서 설명한 시맨틱 멀티미디어 콘텐츠 검색 구조는 풍부한 메타데이터를 바탕으로 잘 구축된 지식 베이스가 전제되어야 한다. 하지만, 모든 정보의 누락 없이 지식 베이스를 구축하기 어려울 뿐만 아니라, 최근 들어 “뿌나(뿌리깊은나무)”, “개콘(개그콘서트)” 등과 같은 네트즌이 만들어낸 프로그램의 줄임말이나, “지랄(뿌리깊은 나무 명대사)”, “띵동(최고의 사랑 명대사)” 등과 같은 이슈성 키워드로 검색을 요청하는 사용자들에게 대응하기 위해 지식 베이스를 업데이트 하기는 어려운 실정이다. 따라서, 본 논문에서는 지식 베이스의 개체 식별 기술을 활용하여 포털 사이트에서 제공하는 연관 검색어를

분석함으로써 사용자 질의를 확장하고 지식 베이스로 구축되지 않은 메타데이터 및 이슈성 키워드에 대한 검색 결과를 제공할 수 있다.

3.1 사용자 질의 해석 결과 조정

사용자 질의 의미 해석은 언제나 사용자의 의도와 다르게 해석할 가능성을 가지고 있다. 따라서 사용자 질의 해석 결과 조정을 통해 시스템이 해석한 최적 해석 결과를 제공함과 동시에 후보 해석 대안을 랭킹하여 제공함으로써 사용자의 의도와 다르게 검색된 해석 결과를 정정하여 검색할 수 있다. 그림 3에서와 같이 “류승완 영화”라고 검색 키워드가 입력되면, 시스템은 지식 베이스 내의 류승완은 감독으로써 연출한 영화 인스턴스가 많기 때문에 “류승완 연출 영화”를 최적 해석 결과로 제공한다. 하지만 사용자의 의도는 “류승완 감독이 출연한 영화”(류승완 감독은 배우로써 출연한 영화가 존재함)를 검색하기 위한 의도로 키워드를 입력하였다고 가정한다면, 시스템에서 제공한 최적 해석 결과는 사용자가 원하는 결과와는 상이하다. 따라서, 시스템에서 제공하는 후보 해석 대안을 통해 키워드의 재입력 없이 검색 결과를 정정하여 원하는 결과를 획득할 수 있게 된다. 그러므로 사용자는 랭킹된 후보 해석 대안들 중에 사용자가 원하는 해석 결과를 선택하고 그 결과를 제공받을 수 있다.

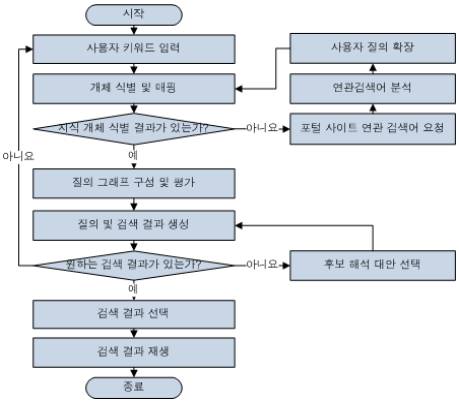


[그림 3] 사용자 질의 해석 결과 조정

3.2 사용자 질의 확장기

제안한 검색 시스템은 지식 베이스로부터 사용자 검색문 의미 해석을 수행함에 따라 지식 베이스에서 식별되지 않는 검색 키워드에 대해서는 질의 그래프를 구성할 수 없으며, 그 검색 결과도 제공할 수 없다. 즉, 지식 베이스로 구축되지 않은 메타데이터 정보를 검색 키워드로 사용할 경우는 질의문 해석 결과도 생성할 수 없으며 검색 결과도 제공되지 않는다. 일반적인 시맨틱 검색 시스템에서도 지식 베이스로 구성되지 않은 키워드에 대한 검색은 그 결과를 제공할 수 없다. 마찬가지로, 키워드 기반의 검색에서도 인덱스로 만들어지지 않은 검색 키워드에 대한 검색 결과는 제공할 수 없다. 하지만, 제안 시스템에서는 사용자 질의 확장기를 통해 입력된 키워드를 확장함으로써 확장된 키워드로 검색을 수행하여 그 결과를 제공할 수 있다.

제안 시스템에서는 입력된 사용자 검색 키워드로부터 개체 식별을 수행하고, 개체 식별 결과가 없을 경우 지식 베이스 내에 사용자 키워드와 일치하는 개체가 없기 때문에 사용자 질의 확장을 수행한다. 사용자 질의 확장 절차는 그림 4에서 보는 바와 같이 개체 식별 결과가 없을 경우, 포털 사이트로부터 연관 검색어를 요청한다.



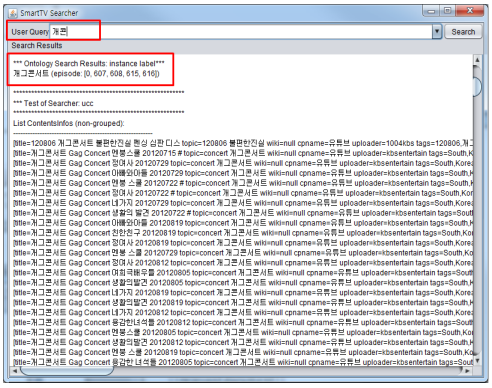
[그림 4] 사용자 질의 확장 절차

2장에서 설명한 바와 같이 제안 시스템은 네이버 연관검색어를 OpenAPI로 요청하고 획득된 연관검색어를 지식 베이스로부터 분석하게 된다. 연관검색어 분석 방법은 우선 획득된 연관검색어를 토큰화하여 개체 식별을 수행한다. 지식 개체로 식별이 된 키워드는 각 키워드에 대해 포털 사이트의 블로그, 카페, 웹문서, 뉴스 등의 관련 문서를 획득하고, 사용자가 초기 입력한 키워드와 획득된 문서 간 TF/IDF 점수를 합산하여 이 중 가장 연관도가 가장 높은 확장 키워드를 추출하게 된다. 확장 키워드 추출을 위한 식은 아래 (1)과 같다. 식 (1)은 포털 사이트로부터 획득된 키워드를 활용하여 수집된 문서에서 사용자로부터 주어진 키워드 k 가 출현한 빈도수를 구하고

이들의 합을 계산한 식이다. tc(k,d)는 문서 d에서 키워드 k의 수이며, $max\{tc(w,d): w \in d\}$는 문서 전체에서 최대 term의 수이다. 즉, 주어진 키워드 k가 문서 내에서 얼마나 자주 등장하는지를 나타내며, $|D|$는 전체 문서의 개수를 표현한다.

$$ck(k,d,D) = \sum_{i=1}^{|D|}\left(\frac{tc(k,d)}{max\ \{tc(w,d): w \in d\}} \times \log\frac{|D|}{|\{d \in D: k \in} \quad (1)$$

그림 5는 사용자 질의 확장기를 통한 검색된 결과를 보여주고 있다. 사용자의 초기 검색 키워드는 “개콘”을 입력하였으며, 질의 확장기를 통해 “개그콘서트”로 키워드를 확장하여 검색된 결과를 제공하고 있다.



[그림 5] 사용자 질의 확장 예

4. 성능 평가 및 비교 분석

4.1 실험환경 구성

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

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documen}{|\{relevant\ document\}|} \quad (2)$$

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documer}{|\{retrieved\ documents\}|} \quad (3)$$

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank} \quad (4)$$

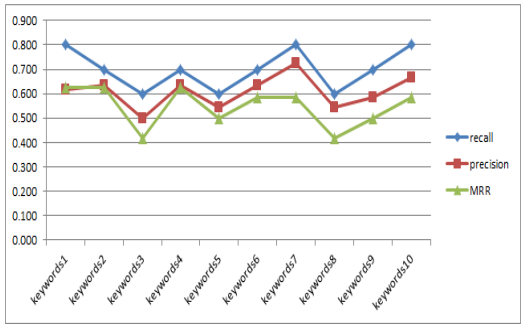
4.2 실험 결과

표 1, 2, 3과 그림 6, 7, 8은 각 시스템에 대하여 검색 예제별로 평가한 재현율, 정확율, MRR을 설명하고 있다. 그리고 표 3과 그림 8은 각 시스템에 대하여 검색 예제별로 평가한 재현율, 정확율, MRR의 평균값을 설명하고 있다. 실험 결과에서 보는 바와 같이 TF/IDF 기반 검색은 다른 시스템에 비하여 재현율은 높으나 정확율이 급격히 감소하고 있음을 알 수 있다. 제안 시스템의 경우 유사한 시맨틱 검색 기법을 사용하는 SPARK와 비슷한 형태의 실험 결과를 보여 주고 있다. 하지만, 검색 성능 면에서 SPARK 시스템보다 조금씩 높은 재현율, 정확율, MRR을 보여주고 있음을 알 수 있다.

논문에서 제안하는 사용자 질의 확장 검색의 경우, 기존 시스템과의 객관적인 성능 평가 및 비교는 불가능하다. 기존 시스템의 경우 인덱스나 지식 베이스에 존재하지 않는 키워드로 메타데이터를 검색하는 경우는 검색 결과가 없는 것이 당연하다. 따라서 타 시스템은 제안 시스템과의 객관적인 성능 평가 및 비교 대상이 되지 못한다.

〈표 1〉 SPARK 실험 결과

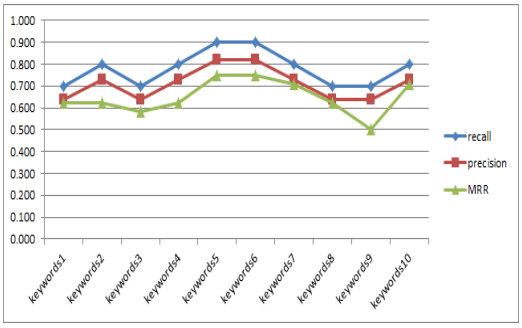
	keyword s1	keyword s2	keyword s3	keyword s4	keyword s5	keyword s6	keyword s7	keyword s8	keyword s9	keyword s10
recall	0.800	0.700	0.600	0.700	0.600	0.700	0.800	0.600	0.700	0.800
precision	0.615	0.636	0.500	0.636	0.545	0.636	0.727	0.545	0.583	0.667
MRR	0.625	0.625	0.417	0.625	0.500	0.583	0.583	0.417	0.500	0.583



[그림 6] 검색 키워드에 대한 실험 결과 (SPARK)

〈표 2〉 제안 시스템 실험 결과

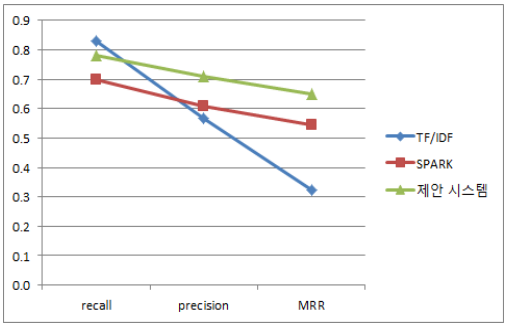
	keyword s1	keyword s2	keyword s3	keyword s4	keyword s5	keyword s6	keyword s7	keyword s8	keyword s9	keyword s10
recall	0.700	0.800	0.700	0.800	0.900	0.900	0.800	0.700	0.700	0.800
precision	0.636	0.727	0.636	0.727	0.818	0.818	0.727	0.636	0.636	0.727
MRR	0.625	0.625	0.583	0.625	0.750	0.750	0.708	0.625	0.500	0.708



[그림 7] 검색 키워드에 대한 실험 결과 (제안 시스템)

〈표 3〉 실험 결과 평균

	recall	precision	MRR
TF/IDF	0.830	0.568	0.324
SPARK	0.700	0.609	0.546
제안 시스템	0.780	0.709	0.650



[그림 8] 실험 결과의 평균

5. 결 론

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. 최근 스마트TV, IPTV와 같은 환경의 한정된 도메인에서 동영상 콘텐츠를 검색하는 시스템으로 활용하면 사용자로부터 높은 만족도를 제공하는 검색 시스템이 될 수 있을 것이다.

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

This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea(NRF) through the Human Resource Training Project for Regional Innovation

참 고 문 헌

[1] 김정민, 정현숙, “방송 온톨로지 구축 및 매칭 기반의 방송 프로그램 검색”, 한국정보기술학회 제9권 제12호, pp.161-171, 2011년 12월.

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

[3] 이경일, “다시 보는 시맨틱 웹 그리고 시맨틱 기술Ⅲ”, White Paper, 솔트룩스, 2010.

[4] 이동균, 권준희, “최근 사용자 관심사를 고려한 소셜 검색 알고리즘”, 한국정보기술학회 제9권 제4호, pp.187-194, 2011년 4월.

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

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

[7] http://semanticwiki.saltlux.com/index.php/의미기반_검색

[8] Naver OpenAPI, http://dev.naver.com/openapi

[9] SPARQL Query Language for RDF (http://www.w3c.org/TR/rdf-sparql-query/), W3C, 2008

[10] SPARQL Query Language for RDF (http://www.w3c.org/TR/rdf-sparql-query/), W3C, 2008.

이 무 훈

・2002년 8월 : 한남대학교 컴퓨터공학과 컴퓨터공학과(공학사)
・2004년 8월 : 한남대학교 컴퓨터공학과(공학석사)
・2008년 10월 ~ 현재 : 한국전자통신연구원 차세대스마트TV연구단 선임연구원
・관심분야 : 시맨틱 검색, 온톨로지, 정보검색, 상황인식 컴퓨팅
・E-Mail : leemh@etri.re.kr

최 의 인

・1982년 : 한남대학교 계산통계학과 (학사)
・1984년 : 홍익대학교 전자계산학과 (석사)
・1995년 : 홍익대학교 전자계산학과 (이학박사)
・1996년 ~ 현재 : 한남대학교 컴퓨터공학과 교수
・2003년 : UCLA 방문 교수
・관심분야 : 시맨틱 웹, 유비쿼터스 컴퓨팅, 모바일, 클라우드 컴퓨팅
・E-Mail : eichoi@hnu.kr