Journal of the Korea Convergence Society
Vol. 11. No. 10, pp. 89-96, 2020

ISSN 2233-4890 / eISSN 2713-6353
https://doi.org/10.15207/JKCS.2020.11.10.089

링크드 오픈 데이터에서 TF-IDF를 이용한 새로운 시맨틱 거리 측정 기법

조정길
성결대학교 컴퓨터공학과 교수

A New Semantic Distance Measurement Method using TF-IDF in Linked Open Data

Jung-Gil Cho
Professor, Department of Computer Engineering, Sungkyul University

요 약 링크드 데이터는 다양한 영역의 데이터세트를 서로 연결할 수 있는 표준 방식의 구조화된 데이터를 가능하게 한다. 그리고 링크드 오픈 데이터(LOD)의 급속한 발전에 따라 연구자들은 시맨틱 유사도 평가와 같은 특정 문제를 해결하기 위해 LOD를 이용하고 있다. 이 논문에서는 LOD-기반 추천 시스템에서 사용될 수 있는 자원 간의 링크드 데이터 시맨틱 거리를 계산하기위한 방법을 제안한다. 이 논문에서 제안된 시맨틱 거리 측정 모델은 LOD-기반 시맨틱 거리와 정보 검색 분야에서 잘 알려진 TF-IDF를 이용한 새로운 링크 가중치를 결합한 유사도 측정을 기반으로 한다. 이 논문의 접근방식의 효과성을 검증하기 위하여 DBpedia와 MovieLens의 혼합 데이터를 사용하여 LOD-기반 추천 시스템의 맥락에서 성능을 평가하였다. 실험 결과는 제안된 방법이 다른 유사한 방법과 비교하여 더 높은 정확도를 나타내었다. 또한 시맨틱 거리 계산의 범위를 넓혀서 추천 시스템의 정확도 향상에 기여하였다.

주제어 : 링크드 오픈 데이터, LOD, 시맨틱 거리, 디비피디아, 자원

Abstract Linked Data allows structured data to be published in a standard way that datasets from various domains can be interlinked. With the rapid evolution of Linked Open Data(LOD), researchers are exploiting it to solve particular problems such as semantic similarity assessment. In this paper, we propose a method, on top of the basic concept of Linked Data Semantic Distance (LDSD), for calculating the Linked Data semantic distance between resources that can be used in the LOD-based recommender system. The semantic distance measurement model proposed in this paper is based on a similarity measurement that combines the LOD-based semantic distance and a new link weight using TF-IDF, which is well known in the field of information retrieval. In order to verify the effectiveness of this paper's approach, performance was evaluated in the context of an LOD-based recommendation system using mixed data of DBpedia and MovieLens. Experimental results show that the proposed method shows higher accuracy compared to other similar methods. In addition, it contributed to the improvement of the accuracy of the recommender system by expanding the range of semantic distance calculation.

Key Words : Linked Open Data, LOD, Semantic Distance, DBpedia, Resource

*Corresponding Author : Jung-Gil Cho(jkcho@sungkyul.ac.kr)
Received August 22, 2020
Revised September 25, 2020
Accepted October 20, 2020
Published October 28, 2020

1. 서론

링크드 데이터는 사유 또는 공개일 수 있으며, 조직과 기업 내부에서 사용하고 비즈니스 파트너와 공유하여 보다 쉬운 통합을 제공하고 상호운용성을 촉진한다. LOD(Linked Open Data)는 개방형 시맨틱 웹 표준과 프리 라이센스를 통해 점점 더 많은 양의 다양한 구조화된 데이터에 액세스 할 수 있는 최근의 커뮤니티 중심 산물이다[1]. LOD 클라우드는 미디어, 지리, 정부, 출판 및 생명 과학과 같은 영역에서 570 개의 데이터세트에 무료로 액세스 할 수가 있다.

최근 몇 년 동안 LOD 클라우드의 인기가 높아지고 있다. LOD의 성공으로 인해 다양한 도메인과 관련된 기계가 이해할 수 있는 형식(주로 RDF[2])으로 웹에서 많은 시맨틱 데이터세트를 자유롭게 사용할 수 있다. LOD는 추천 시스템의 성능을 개선하고 추천 시스템에 내재된 콜드 스타트(cold-start) 문제를 줄이기 위해 추천 시스템에 채택되었다[3,4]. 추천 작업에서 시맨틱 데이터의 사용은 풍부한 데이터 표현을 제공할 뿐만 아니라 다른 도메인에 대해 동일한 접근 방식을 쉽게 채택 할 수 있다. 또한 LOD 데이터세트의 자원이 고유한 URI로 식별되고 의미적으로 서로 연결되어 있기 때문에 동의어(synonymy) 및 다의어(polysemy)와 같은 키워드-기반 접근방식과 관련된 문제가 해결된다[5]. LOD 내의 모든 정보는 도메인 지식이 기본 역할을 하는 LOD 지원 콘텐츠-기반 추천 시스템에서 활용되고 사용될 수 있다. 이와 관련하여 자원 간의 거리를 측정하고 관련성을 식별하는 것은 추천 사항을 제공하기 위한 추천 시스템에 채택 될 수 있으므로 중요한 역할을 한다. 이를 위해 DBpedia와 같은 LOD 데이터세트에서 두 자원 간의 거리와 유사도를 측정하기 위한 다양한 접근법이 제안되었다[6-10].

LOD의 그래프 구조를 사용하는 한 가지 접근방식은 그래프에서 시맨틱 거리로 자원 관련성을 측정하는 것이다. 이 방식은 LOD 그래프에서 서로 연결된 자원이 많을수록 관련성이 높다. 이 개념을 이용한 연구는 자원 관련성 측정의 핵심인 LDSD[6]와 이를 기반으로 한 연구인 자원 유사도(Resim)[7]이다. LDSD 방식은 단일 중간 자원을 통해 직접 연결되거나 간접적으로 연결되는 자원 간의 시맨틱 거리만 계산하기 때문에 두 링크 이상 떨어져있는 자원은 서로 관련이 없는 것으로 간주된다. 예를 들어, Fig. 1은 6 개의 자원이 있는 LOD 데이터세트의 부분 예제이다. LDSD에서는 자원 r_2 및 r_3는 자원 r_1에 도달 할 수 있지만 자원 r_4 및 r_5는 자원 r_1에 도달할 수 없기 때문에 r_1과 관련이 없는 자원으로 측정된다. 반면에 Resim에서는 두 개 이상의 링크에서 떨어져있는 자원에 대한 추가 유사도 측정을 포함하여 자원 r_4 및 r_5가 r_1과 관련이 있는 것으로 측정된다. 그러나 이 방법은 더 멀리 떨어진 링크 거리 자원에 대한 그래프 구조를 고려하지 않고 해당 특성 기준으로만 자원 간의 유사도를 계산하였다[8].

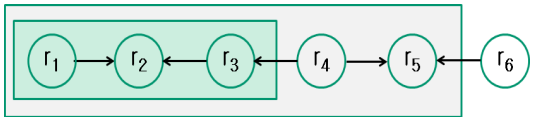

Fig. 1. An example of reachable resources

기존의 LDSD와 Resim 측정에서는 모든 링크를 동일한 가중치를 갖는 것으로 간주하여 시맨틱 거리를 계산하였다. 그러나 이 논문에서는 TF-IDF(Term Frequency-Inverse Document Frequency)[11]를 이용하여 연결된 자원의 네트워크 전체에 걸쳐 각각의 링크에 링크 가중치를 계산하여 시맨틱 거리를 측정한다. 이 논문의 방법은 기존의 시맨틱 거리 접근법의 범위를 하나 이상의 중간 허브를 통해 연결된 자원으로 확장하기 때문에 여러 면에서 유리하다. 때문에 다른 도메인과의 연결이 빈약한 고립된 자원에 대해 훨씬 더 완전한 관련 자원 모음을 만들 수 있다. 특히 LOD-기반 추천 시스템 성능은 희소 자원에 대한 정확도가 떨어지기 때문에 자원 링크 수와 강한 상관관계가 있다. 따라서 LOD 네트워크를 통해 시맨틱 연결을 더 전파하면 자원 범위가 확장되고 더 높은 재현율을 얻을 수가 있다. 또한 잘 연결된 자원의 경우에도 연결을 보다 광범위하게 전파하면 다른 도메인의 관련 자원을 추천 할 수 있다.

2. 관련 연구

Passant(2010)의 연구에서는 LOD 내 자원 사이의 시맨틱 거리를 계산하기 위해 LDSD라는 측정값을 제안했다[6]. 시맨틱 거리 측정에서는 A 자원에서 B 자원으로의 직접 링크와 그 반대로의 링크를 고려했다. 또한 그래프에서 자원 A와 B의 동일한 특성을 통해 같은 인커밍 및 아웃고잉 노드(자원)들을 고려했다. 그러나 두 링크

이상 떨어져있는 자원은 서로 관련이 없는 것으로 간주되었다.

Leal(2012)의 연구에서는 DBpedia에서 자원의 시맨틱 관련성을 계산하는 접근법을 제시했다[12]. 이 접근법에서는 근접성 개념에 기초한 유사도 척도를 제안했는데, 이는 근접성보다는 두 개의 자원이 얼마나 연결되어 있는지를 측정했다. 이는 유사도 측정이 두 노드 사이의 거리와 경로를 모두 고려함을 의미한다. 그러나 LDSD와 마찬가지로 인커밍 노드(자원) 및 자원의 속성은 고려하지 않았다.

Piao(2015)의 연구에서는 자원 유사도(Resim)라는 개선된 링크드 데이터 시맨틱 거리 접근방식을 도입했다[7]. Resim은 자기-유사도, 대칭성 및 최소값과 같은 약점을 극복하면서 기존의 LDSD 접근 방식을 수정했다. 또한 Resim은 데이터세트의 경로 발생에 따라 다른 정규화 방법을 적용하여 Piao(2016)[13]에서 접근방식을 개선했다. 그리고 두 링크보다 더 멀리 떨어진 노드에 대해 속성-기반 유사도 측정을 사용하여 시맨틱 거리에 참여하는 노드 수를 확장했다. 그러나 이 방법은 해당 특성만 기준으로 반영하고 더 멀리 떨어진 링크 거리 자원에 대한 그래프 구조를 고려하지 않고 자원 간의 유사도를 계산하였다.

Alfarhood(2017)연구에서는 LDSD의 시맨틱 거리 접근법의 범위를 넓히는 전파된 링크 데이터 시맨틱 거리(PLDSD)라 불리는 접근법을 소개하였다[8]. PLDSD에서는 잘 알려진 플로이드-워셜(Floyd-Warshall) 알고리즘인 모든 쌍 최단 경로 알고리즘을 사용하여 하나 또는 두 개의 링크 거리에 있는 자원을 넘어 시맨틱 거리 계산을 확장했다. 그러나 이 방법은 LOD 그래프 축소와 시맨틱 거리 전파를 통해 유사도를 측정하였으나 자원에 연결된 링크들에 대해서는 고정된 값으로 자원 간의 유사도를 계산하였다.

3. 시맨틱 거리의 새로운 척도

이 논문에서 제안된 시맨틱 거리 측정 모델은 LOD-기반 시맨틱 거리와 새로운 링크 가중치를 결합한 방법을 기반으로 한다. 그래프에 동일한 가중치를 할당하면 가중치를 정의하지 않은 방법보다 시맨틱 거리를 더 정확하게 나타낼 수가 있으며, 이에 더하여 자원마다 가중치를 다르게 할당하면 동일한 가중치를 할당한 방법보다 더 정확하고 정밀하게 나타낼 수가 있다. 이는 그래프에서 자원들 간의 연결성은 관련성을 나타내기 때문에 자원이 더 많이 연결되고 가중치가 높을수록 더 관련이 있다.

3.1 새로운 척도에 관련된 개념 정의

데이터 수집 단계의 링크드 데이터 그래프에서 링크드 데이터 원칙을 따르는 데이터세트를 나타내기 위해 다음과 같이 정의한다.

정의 1 : LOD 그래프는 방향 그래프 $G=(R, L, T)$ 이다. 여기서 $R=\{r_1, r_2, ..., r_n\}$ 은 URI로 식별되는 자원 집합이고, $L=\{l_1, l_2, ..., l_n\}$ 은 URI로 식별되는 유형이 지정된 링크 집합이며, 그리고 $T=\{t_1, t_2, ..., t_n\}$ 은 자원 쌍을 연결하는 링크의 인스턴스인 트리플의 집합이다. 따라서 RDF 트리플이 〈subject, predicate, object〉 형식의 명령문이므로, $t_i =< r_a, l_j, r_b > \in T$ 는 object $r_b \in R$ 에서 subject $r_a \in R$를 연결하는 predicate $l_j \in L$ 의 인스턴스가 있음을 의미한다.

웹으로 확장하면, LOD 클라우드는 웹에 게시되고 상호 연결된 모든 그래프 G_i의 합집합으로 정의되며 Fig. 2와 같다.

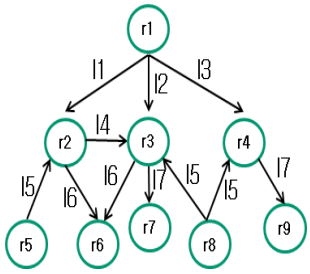

Fig. 2. Example Linked Data Graph

LDSD는 DBpedia와 같은 링크드 데이터 데이터세트에서 두 자원간의 시맨틱 거리를 측정하는 방법이다[6]. LDSD 함수는 다음의 식 1에 정의되어 있으며, 자원 r_a와 r_b 사이에 링크 l_i가 있는지 계산하는 4 개의 함수 $C_d(l_i, r_a, r_b)$로 구성된다.

$$LDSD(r_a, r_b) = \frac{1}{1 + LDSD_{dw}(r_a, r_b) + LDSD_{iw}(r_a, r_b)} \quad (1)$$

여기서 $LDSD_{dw}(r_a, r_b)$는 가중된 직접 거리(C_d)이다.

$$LDSD_{dw} = \sum_i \frac{C_d(l_i, r_a, r_b)}{1+\log(C_d(l_i, r_a, n))} + \sum_i \frac{C_d(l_i, r_b, r_a)}{1+\log(C_d(l_i, r_b, n))}$$

여기서 $LDSD_{iw}(r_a, r_b)$는 가중된 간접 거리(C_{ii}, C_{io})이다.

$$LDSD_{iw} = \sum_i \frac{C_{ii}(l_i, r_a, r_b)}{1+\log(C_{ii}(l_i, r_a, n))} + \sum_i \frac{C_{io}(l_i, r_a, r_b)}{1+\log(C_{io}(l_i, r_a, n))}$$

LDSD 함수는 링크가 있는 경우에 1을 반환하고 그렇지 않으면 0을 반환한다. 자원 n이 있을 때 함수는 자원 r_a 또는 r_b 사이의 다른 모든 자원에 대한 총 링크 수를 계산한다. 식 1의 첫 번째 함수인 $C_d(l_i, r_a, r_b)$는 자원 r_a에서 자원 r_b 로의 직접 링크만 고려한다. 식 1의 두 번째 함수인 $C_d(l_i, r_b, r_a)$는 자원 r_b에서 자원 r_a 로의 역방향 링크인데 다른 링크로 계산된다. 예를 들어, Fig. 2에서 트리플 $t_4 =< r_2, l_4, r_3 >$는 자원 r_2과 r_3 사이의 직접 링크 l_4을 나타낸다. 반대로 자원 r_3에서 r_2 로 링크 l_4이 없기 때문에 역 상황은 존재하지 않는다.

식 1의 세 번째 함수인 $C_{ii}(l_i, r_a, r_b)$는 간접 인커밍 링크를 나타내며 $< r_a, l_i, r_c >$와 $< r_b, l_i, r_c >$를 모두 만족하는 자원 r_c가 있는 경우에만 1을 리턴하고 그렇지 않으면 0을 리턴한다. Fig. 2에서 함수 C_{ii}는 트리플 $t_6 =< r_2, l_6, r_6 >$ 와 $t_7 =< r_3, l_6, r_6 >$로 연결되었다. 이 예에서 r_2과 r_3 사이의 간접 링크는 r_4로 인커밍 l_6 링크를 통해 생성된다.

식 1의 네 번째 함수인 $C_{io}(l_i, r_a, r_b)$는 간접 아웃고잉 링크를 나타내며 $< r_c, l_i, r_a >$ 및 $< r_c, l_i, r_b >$를 모두 만족시키는 자원 r_c가 있는 경우에만 1과 같으며, 그렇지 않은 경우 0이다. Fig. 2에서 C_{io}는 트리플 $t_9 =< r_8, l_5, r_3 >$ 와 $t_{10} =< r_8, l_5, r_4 >$로 연결되었다. 이 예에서 r_3와 r_4 사이의 간접 링크는 r_8에서 아웃고잉 l_5 링크를 통해 생성된다.

3.2 새로운 시맨틱 거리 척도

기존의 측정방법인 LDSD, Resim, PLDSD에서는 모든 링크 l_x를 동일한 가중치를 갖는 것으로 간주하여 시맨틱 거리를 계산했다. 그러나 이 논문에서는 각각의 링크에 링크 가중치를 계산하기 위해 정보 검색과 데이터 마이닝에서 이용하는 가중치인 TF-IDF(Term Frequency-Inverse Document Frequency)[11]를 사용한다. TF-IDF는 문서들 그룹 내에서 문서에 있는 단어(term)의 중요성을 평가하는 데 사용된다. 이 논문의 범위에서 자원에 대한 링크의 중요성은 LOD 데이터세트에 있는 자원들 및 링크들 전체 컬렉션을 기반으로 평가된다. 따라서 TF-IDF의 단어 빈도($tf(d,t)$)와 역문서 빈도($idf(d,t)$)는 각각 링크 빈도($lf(l_x, r_a, r_b)$)와 역자원 빈도($irf(l_x, r_a, r_b)$)로 변형하여 사용한다.

이 논문에서의 링크 가중치는 동적으로 계산되어 링크와 링크된 자원 간의 관계 외에도 링크와 데이터세트의 다른 링크 간의 관계도 고려한다. 이 방법은 문서들 그룹 자원에 있는 많은 양의 링크에 대하여 각각의 링크 가중치를 계산하기 위해 전체 LOD 데이터세트를 순회해야한다는 것이다. 그래서 링크 가중치 값은 전처리 단계에서 한 번 계산되어 저장한 다음에 LOD 엔진에 통합하여 필요할 때 즉시 가중치를 사용할 수 있다. 또한 이 논문에서는 링크 가중치 계산을 위해 모든 링크 정보가 필요하기 때문에 링크인 W_{l_i} 대신에 $W(l_x, r_a, r_b)$로 나타낸다. 그리고 우리의 접근방식은 제안된 링크 가중치 범위인 [0-1]의 제약조건을 충족하지 않는 가중치를 생성하기 때문에 이러한 값을 [0-1] 범위로 다시 조정해야한다. TF-IDF 단어 가중치에 대응된 초기 비척도 링크 가중치 $W_{ns}(l_x, r_a, r_b)$는 다음의 식 2와 같이 계산한다.

$$W_{ns}(l_x, r_a, r_b) = lf(l_x, r_a, r_b) \times irf(l_x, r_a, r_b) \quad (2)$$

이 식에서, 링크 빈도 $lf(l_x, r_a, r_b)$는 자원 r_a 또는 r_b를 다른 자원으로 연결하는 링크 l_x의 평균 정규화 빈도이다. 링크 빈도는 다음의 식 3과 같이 자원 r_a 또는 r_b에 대한 인커밍 및 아웃고잉 링크 총 수로써 정규화된 자원 r_a 또는 r_b에 대한 속성 l_x의 인커밍 및 아웃고잉 링크 총 수로 계산한다.

역자원 빈도 $irf(l_x, r_a, r_b)$는 다음의 식 4와 같이 링크 l_x의 총 인스턴스로 나눈 LOD 데이터세트의 총 자원 수이다.

$$lf(l_x, r_a, r_b) = \frac{(\frac{\sum_j C_d(l_x, r_a, r_j) + \sum_j C_d(l_x, r_j, r_a)}{\sum_i \sum_j C_d(l_i, r_a, r_j) + \sum_i \sum_j C_d(l_i, r_j, r_a)}) + (\frac{\sum_j C_d(l_x, r_b, r_j) + \sum_j C_d(l_x, r_j, r_b)}{\sum_i \sum_j C_d(l_i, r_b, r_j) + \sum_i \sum_j C_d(l_i, r_j, r_b)})}{2} \quad (3)$$

$$WLDSD(r_a, r_b) = \frac{1}{1 + WDC(r_a, r_b) + WDC(r_b, r_a) + WTIC_{ii}(r_a, r_b) + WRIC_{io}(r_a, r_b)} \quad (6)$$

$$irf(l_x, r_a, r_b) = \log(\frac{\sum_i r_i}{\sum_i \sum_j C_d(l_x, r_i, r_j)}) \quad (4)$$

마지막으로 우리의 접근방법을 사용하여 계산된 링크 가중치는 [0-1] 범위의 제약조건을 충족하지 않기 때문에 다음의 식 5와 같이 [0-1] 범위에서 다시 척도 되어야 한다.

$$W(l_x, r_a, r_b) = \frac{W_{ns}(l_x, r_a, r_b) - \min}{\max - \min} \quad (5)$$

이 식에서 min은 최소 계산된 링크 가중치 값이고, max는 최대 계산된 링크 가중치 값이다.

이 논문의 시맨틱 거리 측정 식은 링크 가중치 함수인 $W(l_x, r_a, r_b)$를 기존 LDSD 식에 추가하여 정의한다. $W(l_x, r_a, r_b)$는 자원 r_a와 r_b에 대하여 각각의 링크 가중치를 나타내므로 식 6에 도시된 바와 같이 식 1에 있는 모든 $C_d(l_x, r_a, r_b)$ 함수에 곱한다.

여기서 $WDC(r_a, r_b)$는 $W(l_x, r_a, r_b)$에 의해 가중된 직접 거리(C_d)이다. 속성이 l_x 인 각각의 링크의 경우 다음과 같다.

$$WDC(r_a, r_b) = \sum_x (\frac{C_d(l_x, r_a, r_b)}{1 + \log(C_d(l_x, r_a, n))} \times W(l_x, r_a, r_b))$$

$WDC(r_b, r_a)$는 $W(l_x, r_a, r_b)$에 의해 가중된 직접 거리(C_d)이다. 속성이 l_x 인 각각의 링크의 경우 다음과 같다.

$$WDC(r_b, r_a) = \sum_x (\frac{C_d(l_x, r_b, r_a)}{1 + \log(C_d(l_x, r_b, n))} \times W(l_x, r_a, r_b))$$

$WTIC_{ii}(r_a, r_b)$는 자원 r_a에 대한 모든 인커밍 무형 간접 링크의 로그에 의해 정규화되고, 다음과 같이 $W(l_x, r_a, r_b)$에 의해 가중화되는 자원 r_a와 r_b 사이의 인커밍 무형 간접 거리 (TIC_{ii})이다.

$$WTIC_{ii}(r_b, r_a) = \sum_x (\frac{C_{ii}(l_x, r_a, r_b)}{1 + \log(C_{ii}(l_x, r_a, n))} \times W(l_x, r_a, r_b))$$

$WTIC_{io}(r_a, r_b)$는 자원 r_a에 대한 모든 아웃고잉 무형 간접 링크의 로그에 의해 정규화되고, 다음과 같이 $W(l_x, r_a, r_b)$에 의해 가중화되는 자원 r_a와 r_b 사이의 아웃고잉 무형 간접 거리 (TIC_{io})이다.

$$WTIC_{io}(r_b, r_a) = \sum_x (\frac{C_{io}(l_x, r_a, r_b)}{1 + \log(C_{io}(l_x, r_a, n))} \times W(l_x, r_a, r_b))$$

모든 링크 가중치 $W(l_x, r_a, r_b)$의 값은 0과 1 사이의 양의 유리수이며, ($0 \le W(l_x, r_a, r_b) \le 1$) 이다. 링크 가중치 기반 요소 $W(l_x, r_a, r_b)$는 모든 링크 기반 작업에 도입된다. 따라서 링크 가중치가 높은 링크에 대해 더 높은 직접 및 간접 거리 값이 생성된다. 반대로 링크 가중치가 낮으면 이러한 링크에 대한 중요도가 덜하며, 해당 가중치가 0인 경우에는 일부 링크 특성이 취소된다.

4. 실험 결과와 성능 평가

4.1 실험 결과

이 논문에서 제안된 시맨틱 거리 측정은 추천 알고리즘으로 사용 시에 성능 측면에서 평가한다. 이를 위하여 실험 평가에 사용된 데이터베이스는 Movielens 1M

Dataset[14]이며, 4000편의 영화에서 6000명의 사용자로부터 백만 개의 평가를 받았다. 또한 이를 바탕으로 사용자 모델인 영화에 대한 사용자, 영화 및 사용자 등급을 저장할 수 있는 하나의 데이터베이스를 모델링하고 영화의 내용도 저장했다. 그리고 MovieLens 1M의 각각의 영화에 대해 RDF 식별자(URI)를 제공하는 Mapping Movielens2DBpedia[15]를 사용하여 DBpedia 파일에 매핑된 MovieLens를 데이터베이스로 가져 왔다. 따라서 이 초기 설정을 통해 온라인으로 SPARQL 쿼리를 통해 DBpedia에서 자원 및 모든 연결에 액세스 할 수 있다. 알고리즘이 실행되는 동안 사용자마다 링크 가중치 및 유사도 측정 결과를 저장한다. 각각의 사용자와 자원의 유사도는 다음의 식 7과 같이 계산한다.

$$similarity(u_i, r_a) = \frac{\sum_{r_b \in Profile(u_i)} (1 - SemantidDistance(r_a, r_b))}{|Profile(u_i)|} \quad (7)$$

여기서 $SemanticDistance(r_a, r_b)$는 이 논문을 포함하여 평가에 사용된 시맨틱 거리 방법들이다. $Profile(u_i) = \{r_1, r_2, ..., r_m\}$는 사용자가 이전에 즐겨 찾았던 자원 집합인 사용자 프로필이다.

자원의 결과 목록을 사용자별 내림차순으로 정렬한 다음에, 이 테스트 자원을 각각의 시맨틱 거리 접근방식의 효과를 측정하기 위해 사용하였다. 표준 평가 척도인 F_1-score 및 MRR(Mean Reciprocal Rank)을 평가 방법으로 사용하였으며, 정밀도와 재현율의 조화 평균인 F_1-score는 다음과 같이 계산한다[16,17].

$$F_1 - score = 2 \times \frac{precision \times recall}{precision + recall}$$

F_1-score는 0.0~1.0 사이의 값을 가지며 높은 값일수록 평가 척도가 좋으며, 정밀도와 재현율의 둘 중 하나라도 0이 되면 F_1-score도 0이 된다. 또한 MRR은 사용자와 관련된 첫 번째 항목의 평균 순위를 나타내며 다음과 같이 계산한다.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

여기서 $rank_i$는 쿼리 Q_i에서 관련 결과의 최고 순위이다.

Table 1은 실험 데이터세트를 사용하여 F_1-score 및 MRR을 평가한 실험 결과이다. F_1-score 값은 상이한 순위 결과 컷오프인 5, 10, 20으로 제시했다. 이 논문의 평가 결과는 F_1@5, F_1@10, F_1@20, MRR 값이 각각 0.055, 0.064, 0.062, 0.048의 성능을 보여주었으며, 이러한 결과로 이 논문의 모델 성능이 우수함을 입증하였다.

Table 1. The result of performance comparison with other methods

	F1@5	F1@10	F1@20	MRR
LDSD	0.036	0.045	0.047	0.029
Resim	0.046	0.056	0.055	0.039
PLDSD	0.051	0.061	0.061	0.045
this paper	0.053	0.064	0.065	0.048

4.2 비교 평가

Table 1의 결과에서 알 수가 있듯이, 이 논문의 접근방법 정확도는 모든 지표(F_1-score 및 MRR)에서 다른 접근방법인 LDSD, Resim, PLDSD을 모두 능가했다. F_1-score는 또한 상위 5개 결과(F_1@5)에 대하여 이 논문의 경우 0.053의 점수, PLDSD의 경우 0.051의 점수, LDSD의 경우 0.036의 점수, Resim의 경우 0.046 인 결과를 도출하였다. 이러한 결과는 Fig. 3에 표시된 다른 결과 컷오프 지점인 F_1@10 와 F_1@20에서도 유지되었다. F_1@10의 결과는 이 논문, PLDSD, Resim, LDSD가 각각 0.064 0.061 0.056 0.045인 점수를 도출하였으며, F_1@20의 결과는 이 논문, PLDSD, Resim, LDSD가 각각 0.065 0.061 0.055 0.047인 점수를 도출하였다.

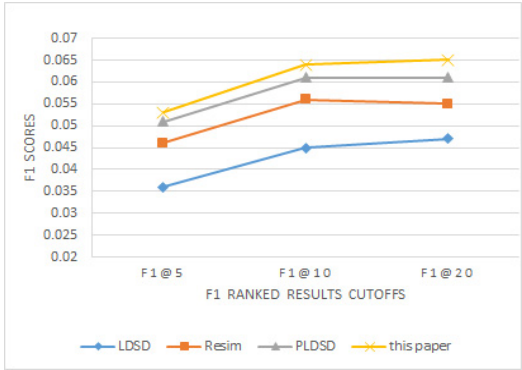

Fig. 3. F_1-scores at different ranked results cutoffs for other methods

Fig. 4에서와 같이 MRR 값은 이 논문의 성능이 LDSD에 비해 65.5 % (0.048 대 0.029)가 개선되고, Resim에 비해 23.1 % (0.048 대 0.039)가 개선되었으며, PLDSD에 비하여 6.7 %(0.048 대 0.045)가 개선되었음을 보여주었다.

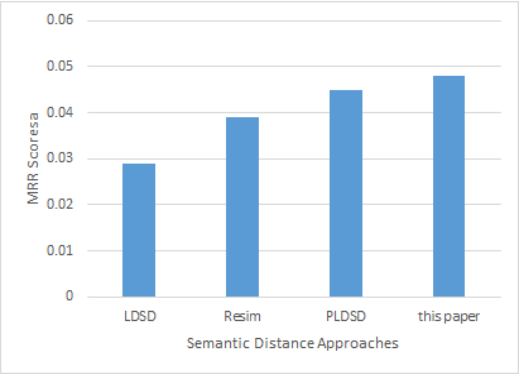

Fig. 4. Comparison of MRR scores with other methods

5. 결론

이 논문에서는 링크드 데이터의 자원에 대한 시맨틱 거리를 측정하기 위해 새로운 접근방식을 제안했다. 우리의 접근 방식에서는 추가 자원을 포함하도록 기존의 시맨틱 거리 접근방식의 범위를 확장하였는데, 자원에 대한 링크의 중요성은 LOD 데이터세트에 있는 자원들 및 링크들 전체 컬렉션을 기반으로 평가되었다. 우리는 두 링크 이상으로 떨어진 자원을 기반으로 시맨틱 거리를 효율적으로 계산하기 위해 TF-IDE를 변형하여 사용하였다. 평가 결과는 링크 가중치를 사용하지 않는 기존 방법들을 능가하는 성능을 보였다. 이 결과는 그래프에서 자원을 더 멀리 고려하여 시맨틱 거리 계산을 하면 LOD 기반 추천 시스템의 정확도가 향상됨을 보여준다. 향후 연구로는 확장된 시맨틱 거리 전파가 다른 도메인 영역에 미치는 영향을 분석할 계획이다.

REFERENCES

[1] C. Bizer, T. Heath & T. Berners-Lee. (2009). Linked Data-The Story So Far. *International Journal on Semantic Web and Information Systems, 5(3),* 1-22. DOI : 10.4018/jswis.2009081901

[2] Google. (2004). RDF vocabulary description language 1.0: RDF schema. W3C[Online]. https://www.w3.org/2001/sw/RDFCore/Schema/20021 2bwm/

[3] V. C. Ostuni, T. D. Noia, E. D. Sciascio & R. Mirizzi. (2013). Top-n recommendations from implicit feedback leveraging linked open data. *In Proceedings of the 7th ACM conference on Recommender systems,* 85-92. DOI : 10.1145/2507157.2507172

[4] A. Passant. (2010). dbrec: Music Recommendations Using DBpedia. *In ISWC 2010 SE-14,* 209-224. DOI : 10.1007/978-3-642-17749-1_14

[5] S. E. Middleton, D. De Roure & N. R. Shadbolt. (2009). *Ontology-based recommender systems.* In Handbook on ontologies, 779-796.

[6] A. Passant. (2010, March). Measuring Semantic Distance on Linking Data and Using it for Resources Recommendations. *In AAAI Spring Symposium: Linked Data Meets Artificial Intelligence* (Vol. 77, p. 123).

[7] G. Piao, S. S. Ara & J. G. Breslin, (2015). Computing the Semantic Similarity of Resources in DBpedia for Recommendation Purposes. *In 5th Joint International Semantic Technology Conference.* (pp. 185-200). Springer, Cham. DOI: 10.1007/978-3-319-31676-5

[8] S. Alfarhood, K. Labille & S. Gauch. (2017) PLDSD: Propagated Linked Data Semantic Distance. *IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises(WETICE),* 278-283. DOI: 10.1109/WETICE.2017.16

[9] G. O. Silva, F. A. Durao & M. Capretz, (2019). PLDSD: Personalized Linked Data Semantic Distance for LOD-Based Recommender Systems. *iiWAS2019.* DOI: 10.1145/3366030.3306041

[10] S. Alfarhood, S. Gauch & K. Labille. (2019). Semantic Distance Spreading Accross Entities in Linked Open Data. *Information 2019, 10(15),* 1-15. DOI: 10.3390/info10010015

[11] D. S. Park & H. J. Kim. (2018). A Proposal of Join Vector for Semantic Factor Reflection in TF-IDF Based Keyword Extraction. *Journal of KIIT, 16(2),* 1-16. DOI : 10.14801/JKIIT.2018.16.2.1

[12] J. P. Leal, V. Rodrigues & R. Queir´os. (2012). Computing semantic relatedness using dbpedia. *Symposium on Languages, Applications and Technologies, 1st* (pp. 133-147). Schloss Dagstuhl. DOI: 10.4230/OASIcs.SLATE.2012.133

[13] G. Piao & J. G. Breslin. (2016). Measuring Semantic Distance for Linked Open Data-enabled Recommander Systems. *SAC '16: Proceedings of the 31st Annual ACM Symposium on Applied Computing,* 315-320. DOI: 10.1145/2851613.2851839

[14] Google. (2020). *Movielens 1M Dataset*. grouplens [Online]. https://grouplens.org/datasets/movielens/1m/

[15] Google. (2020). MappingMovielens2DBpedia. researchGate [Online]. https://www.researchgate.net/publication/297369577_mapping-movielens-dbpedia

[16] J. G. Cho. (2020). A location localization method using Smartphone sensor on a subway. *Journal of the Korea Convergence Society, 11(3),* 37-43. DOI : 10.15207/JKCS.2020.11.3.037

[17] D. Khongorzul, S. M. Lee & M. H. Kim. (2019). OrdinalEncoder based DNN for Natural Gas Leak Prediction. *Journal of the Korea Convergence Society, 10(10),* 7-13. DOI : 10.15207/JKCS.2019.10.10.007

조 정 길(Jung-Gil Cho) [정회원]

· 1987년 2월 : 숭실대학교 전자계산학과(공학사)
· 1993년 2월 : 숭실대학교 정보과학대학원(이학석사)
· 2003년 2월 : 충북대학교 전산과(이학박사)
· 2004년 3월 ~ 현재 : 성결대학교 컴퓨터공학과 교수
· 관심분야 : XML 문서관리, 정보 검색, 스마트폰 사용성
· E-Mail : jkcho@sungkyul.ac.kr