wikidata를 이용하는 상황 인지 엔티티 링크 프레임워크

장서윤O, 박종현*
OO충남대학교 컴퓨터융합학부,
*한신대학교 SW역량강화센터
e-mail: pineai@cnu.ac.krO, jh7park@hs.ac.kr*

Context-aware entity link framework using wikidata

SeoYoon JangO, Jong-Hyun Park*
ODepartment of Computer Science & Engineering Chungnam National University,
*Hanshin University, SW Competency Enhancement Center

● 요 약 ●

사용자의 관심사를 고려하면 상황 인지 서비스의 질을 높일 수 있다. 기존의 사용자의 관심사를 고려하는 서비스에는 지식베이스(KB)가 사용 되었으나, 최근 새로운 방법인 wikidata를 이용한 엔티티 링크를 활용한 방법도 활발히 연구가 진행되고 있다. wikidata가 적용된 엔티티 링크는 기존의 KB를 이용하는 방법보다 데이터의 변경, 보완이 쉽고 가볍다. 이에 본 논문에서는 wikidata가 적용된 엔티티 링크 모델을 이용한 상황인지 서비스를 제공 할 수 있는 프레임워크를 제안한다.

키워드: wikidata, 엔티티 링크 모델(entity linker model), 상황 인지 프레임워크(context-aware framework)

I. Introduction

최근 사용자에게 적합한 상황 인지 서비스를 제공하기 위하여 사용자의 관심사를 고려하는 연구가 활발히 진행 되고 있다. 사용자의 소셜 미디어 데이터를 분석하면 사용자의 최근 개인적 관심사나 사회적 관심사를 이해 할 수 있다[1,2,3]. 소셜 미디어 데이터 분석에는 여러 기법이 있으나, 사용자 관심사를 찾기 위한 사용자 프로파일링 연구가 활발히 진행 되고 있다. 사용자가 작성하거나 선호하는 문장에서 연결성을 찾는 연구는 주로 지식 베이스(KB, Knowledge Bases)를 이용한다. KB를 이용한 연구는 대부분 DBpedia[4] 또는 YAGO[5]와 같은 위키피디아에서 파생된 지식을 이용하였다. 이 KB들은 위키피디아 페이지의 정보 및 범주에서 정보를 수집하여 자동으로 관리되어 편집이 불가능하다. 하지만 wikidata[6]는 편집이 가능한 다국어 지식 데이터베이스이다. 또한, wikidata는 구조화 된 데이터로만 입력 할 수 있어 지식 그래프(Knowledge Graph)를 구성하기 용이하다. 지식 그래프는 최근 활발히 연구 되고 있는 개체 연결에 대한 새로운 접근 방식이다.[7,8,9] 구조화 된 wikidata의 데이터 모델은 지식 그래프 및 새로운 학습법의 도입으로 최근에 기업 정보 연결을 위한 데이터베이스로 인기를 얻고 있다.

한편, 기존에 위치 정보와 소셜 미디어 데이터를 이용하여 사용자의 관심사에 맞는 서비스를 제공하는 연구가 있었다[10,11]. 기존 연구들 중 하나[11]는 KB의 일종인 위키피디아의 범주를 이용하여 주요 키워드를 추출하였다. 추출된 키워드는 유사도를 분석하여 사용자에게 관심항목을 제안 한다. 하지만, 기존의 연구는 위키피디아의 범주만을 이용하여 지식 그래프를 적용하기 어렵다.

본 연구에서는 wikidata의 엔티티 연결을 이용하여 사용자 관심사를 고려하는 상황인지 서비스를 할 수 있는 엔티티 링크 프레임워크를 제안한다. 기존의 상황인지 프레임워크보다 wikidata와 연결 수정 및 개선이 용이하고, 연결성 분석이 가벼운 특징을 가진다.

II. Related works

wikidata[6]는 wikimedia에서 제공하는 위키(wiki, 불특정 다수가 협업을 통해 직접 내용과 구조를 수정할 수 있는 웹사이트)지만, 구조화된 위키 페이지를 저장하는 특징을 가진다. 항목(item)은 Q-id로 식별 되며 여러 데이터 필드로 구성된다. 레이블은 엔티티의 기본 이름을 저장한다. 이것은 명칭을 구분하기 위해 항목을 설명하며, 별칭을 가진다. 항목은 RDF 스타일의 문장을 저장하며 참조에 의해 연결 될 수 있고 한정어(qualifiers)로 보다 정밀하게 만들어질 수 있다. 또한, 항목은 다른 wikimedia 프로젝트인 wikipedia등의 엔티티에 해당하는 페이지에 연결 되는 사이트 링크를 가질 수 있다.

예를 들어, 스타벅스(Starbucks, Q37158)인 경우 Fig. 1과 같은 페이지와 정보를 가진다. wikidata를 이용하여 엔티티 링크를 연결하는 openTapioca[12]는 실시간으로 훈련하고 실행하며 동기화를 지원한다. 이 연구는 wikidata만을 이용하여 Named Entity Linking을 하는 프레임워크를 제안하는 것을 목적으로 하였으며 가볍고 효과적인 추천 결과를 도출하였다. 하지만, 기존 연구는 사람, 위치 및 조직에 관련된 내용만을 연결 하도록 제한하였다.

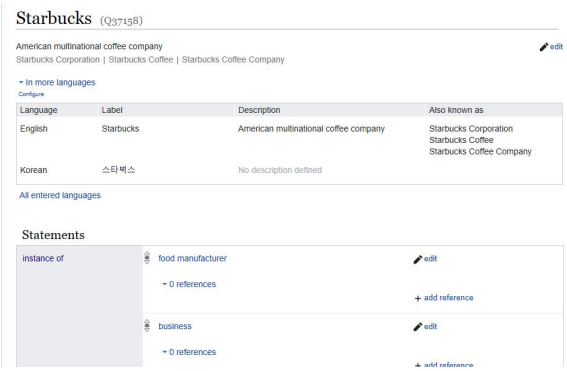

Fig. 1. wikidata의 starbucks 페이지

기존에 KB를 이용하는 사용자의 관심에 맞는 상황인지 지능형 시스템은 지속적으로 연구 되어 왔다.[4, 5,11] 기존의 연구들은 wikipedia의 범주를 이용하여 소셜 미디어에서 주요 단어를 추출하여 사용자의 관심을 기반으로 상황에 맞는 서비스를 제공한다. 기존의 방법은 지식 그래프를 적용하기 어려우며, 주요 단어 간 연결 시스템이 복잡하여 연결 시스템 수정이나 개선에 상당한 지연이 발생한다. 하지만, wikidata를 이용하는 엔티티 연결은 실시간으로 동기화가 가능하며 기본 데이터에 직접 개입이 가능하고 연결 수정 및 개선이 용이하다. 이에 본 연구에서는 wikidata의 엔티티 링크 모델이 적용된 상황인지 서비스 엔티티 링크 프레임워크를 제안한다.

III. wikidata를 이용하는 상황 인지 엔티티 링크 프레임워크

Fig. 2는 wikidata를 이용하는 사용자의 관심사를 고려하도록 설계된 상황 인지 엔티티 링크 프레임워크이다.

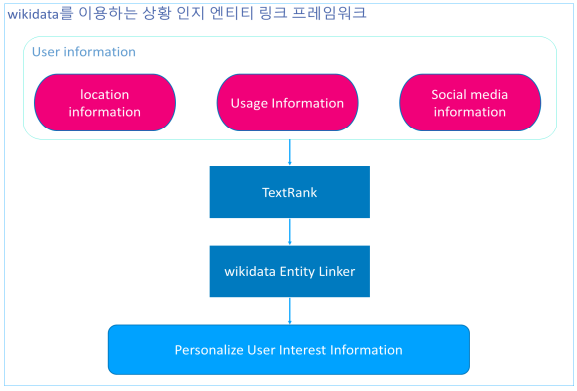

Fig. 2. wikidata를 이용하는 상황 인지 엔티티 링크 프레임워크

기존 연구와의 비교를 위해 우리는 wikipedia 데이터는 사용하지 않는다. wikidata의 엔티티 링커는 상황인지 서비스에 사용 할 수 있도록 openTapioca를 일부 수정하였다. 프레임워크는 wikidata에서 제공하는 정보를 동기화 하여 연결 정보를 수정하고 개선 할 수 있도록 고안되었다.

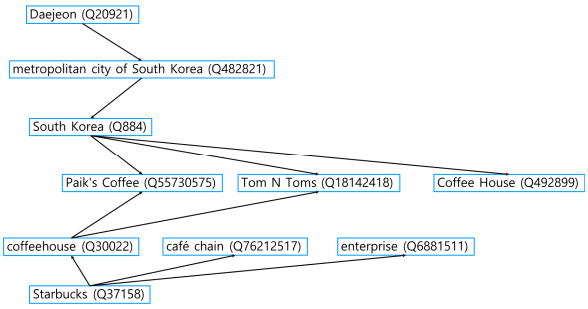

Fig. 3. wikidata로 생성된 엔티티 연결 그래프

사용자의 관심사를 찾기 위한 초기 인덱스는 기기의 검색기록, 소셜 네트워크 데이터 그리고 위치 기록을 이용하였다. 인덱스의 추출은 TextRank[13]를 이용하였다. 또한, 너무 넓은 범위의 엔티티 연결을 막기 위해 초기 인덱스 항목을 태깅하여 기존 분류법의 범위를 기준으로 최대 5단계 이상의 연결을 넘어가지 않도록 관리한다. 예를 들어, 최근의 초기 인덱스가 소셜 미디어 데이터에서 스타벅스(Starbucks, Q37158)와 상황인지 위치 정보에서 대전(Daejeon, Q20921)을 얻었다고 가정하면, 엔티티 연결은 Fig. 3과 같은 구조를 보인다. 이 엔티티 연결을 이용하여 상황인지 서비스는 사용자에게 Paik's Coffee (Q55730575), Tom N Toms (Q18142418)등을 추천한다.

IV. Conclusions

본 논문에서는 wikidata의 entity link를 이용하는 상황 인지 프레임워크를 제안하였다. 기존의 wikipedia등의 KB 정보를 이용하여 사용자의 관심사에 맞는 상황인지 서비스를 제공하는 프레임워크들과는 다르게, wikidata의 entity link를 이용하면 데이터의 연결성 변경이 쉽고 실시간으로 wikidata와 동기화하여 새로운 정보를 적용 할 때의 비용 부담이 적다. 하지만, wikidata의 entity link만 사용할 경우 관련 링크 혹은 문서로 데이터가 제공 될 경우, 별도로 긴 문장을 entity로 분석 하는 기능이 없어 정보로서 가공하기 어려운 문제가 발생한다. 향후, 이러한 문제점을 해결하기 위하여 기존의 wikipedia를 이용하여 문서 유사도 측정을 하는 기능을 추가한 프레임워크를 연구할 예정이다.

ACKNOWLEDGEMENT

이 성과는 2018년도 과학기술정보통신부의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2018R1A2B6008965).

REFERENCES

[1] Seghouani, Nacéra Bennacer, Coriane Nana Jipmo, and Gianluca Quercini. "Determining the interests of social media users: two approaches." Information Retrieval Journal 22.1-2 (2019): 129-158.

[2] Spasojevic, Nemanja, et al. "Lasta: Large scale topic assignment on multiple social networks." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014.

[3] Pennacchiotti, M., Silvestri, F., Vahabi, H., & Venturini, R. (2012). Making your interests follow you on twitter. In CIKM, pp. 165–174.

[4] S¨oren Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. The semantic web, pages 722–735.

[5] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In Proceedings of the 16th International Conference on World Wide Web, pages 697–706. ACM.

[6] Denny Vrande˘ci´c and Markus Kr ¨otzsch. 2014. Wikidata: A free collaborative knowledge base. Communications of the ACM.

[7] Albert Weichselbraun, Philipp Kuntschik, and Adrian M.P. Bras¸oveanu. 2018. Mining and Leveraging Background Knowledge for Improving Named In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics,WIMS ’18, pages 27:1–27:11, New York, NY, USA. ACM.

[8] Daniil Sorokin and Iryna Gurevych. 2018. Mixing Context Granularities for Improved Entity Linking on Question arXiv:1804.08460 [cs].

[9] Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-End Neural Entity Linking. arXiv:1808.07699 [cs].

[10] Bao, Jie, et al. "Recommendations in location-based social networks: a survey." GeoInformatica 19.3 (2015): 525-565.

[11] SeoYoon,Jang. JiHoon,Kang, “Location Information Sharing Objects (LISO)”. in Progress.

[12] Lu et al. 2012 - Twitter User Modeling and Tweets Recommendation Based on Wikipedia Concept Graph

[13] Barrios, Federico, et al. "Variations of the similarity function of textrank for automated summarization." arXiv preprint arXiv:1602.03606 (2016).