Towards Building a Knowledge Graph with Open Data – A Roadmap

Farouk Musa Aliyu[1(✉)] and Adegboyega Ojo[2]

[1] Federal University Birnin Kebbi, Birnin Kebbi, Kebbi, Nigeria
musa.farouk@fubk.edu.ng
[2] Insight Centre for Data Analytics, National University of Ireland, Galway (NUIG), Galway, Ireland
adegboyega.ojo@insight-centre.org

Abstract. With the increasing interest in knowledge graph over the years, several approaches have been proposed for building knowledge graphs. Most of the recent approaches involve using semi-structured sources such as Wikipedia or information crawled from the web using a combination of extraction methods and Natural Language Processing (NLP) techniques. In most cases, these approaches tend to make a compromise between accuracy and completeness. In our ongoing work, we examine a technique for building a knowledge graph over the increasing volume of open data published on the web. The rationale for this is two-fold. First, we intend to provide a foundation for making existing open datasets searchable through keywords similar to how information is sought on the web. The second reason is to generate logically consistent facts from usually inaccurate and inconsistent open datasets. Our approach to knowledge graph development will compute the confidence score of every relationship elicited from underpinning open data in the knowledge graph. Our method will also provide a scheme for extending coverage of a knowledge graph by predicting new relationships that are not in the knowledge graph. In our opinion, our work has major implications for truly opening up access to the hitherto untapped value in open datasets not directly accessible on the World Wide Web today.

Keywords: Knowledge graph · Open data

1 Introduction

In this section, we briefly introduced knowledge graph and open data.

1.1 Open Data

According to open definition [7] open data refers to data that "anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)." From this definition, open data includes any kind of data that can be freely accessed, modified and share on the web. Open data exist in different formats including text documents, spreadsheet, structured documents in RDF or JSON format, pictures, geographic files formats, etc. Popular examples of common open data sets include those published in government portals such as data.gov.* (e.g. uk, i.e., and

V. Odumuyiwa et al. (Eds.): AFRICOMM 2017, LNICST 250, pp. 157–162, 2018.
https://doi.org/10.1007/978-3-319-98827-6_13

es). Examples of open data portals in Africa include http://data.edostate.gov.ng of the Edo State Government in Nigeria, http://www.opendata.go.ke/ of the Kenyan Government and http://dataportal.opendataforafrica.org/ maintained by the African Development Bank. See Fig. 1 for example of an open data portal. Related to open data are also public data and resource such as DBpedia [14], YAGO [3], Geonames[1], Wikipedia, word-Net[2], dbtune.org, New York Times dataset[3], opendatacommunities.org datasets, etc. Open data covers a wide range of domains which are heterogeneous in nature and noisy. Open data, therefore, reveal a large variation in quality. Applications consuming this data need to therefore, engage in other processing steps to deal with the inconsistencies and misleading information. The issues with open data include: accuracy, representation, integration and linking. One way to address this problem is by integrating islands of non-consistent open datasets to build a more consistent global dataset in the form of knowledge graph.

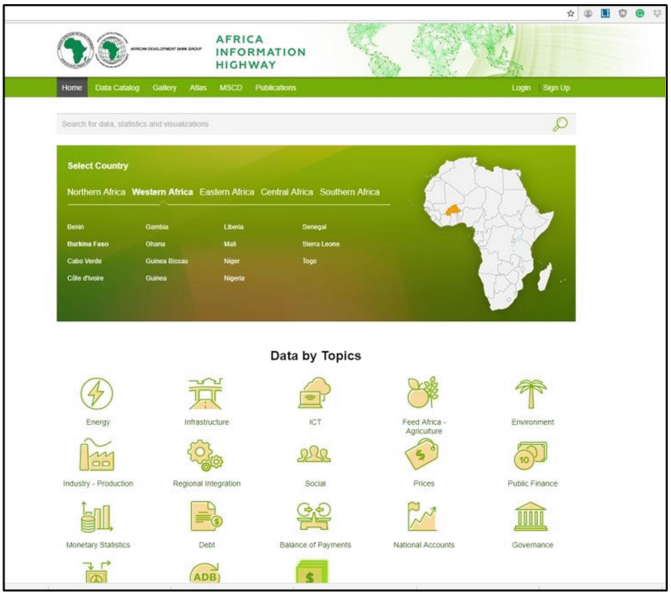

Fig. 1. Example of an open data portal (http://dataportal.opendataforafrica.org/)

1.2 Knowledge Graph

There is no generally agreed definition of what a knowledge graph is. The term knowledge graph was originally used by Google when introducing their knowledge graph [5] in 2012. Ever since, researchers have often used the term to refer to semantic

[1] www.geonames.org/.

[2] https://wordnet.princeton.edu/.

[3] www.nytimes.com/.

web repositories such as DBpedia [14] and YAGO [3]. [4] Defines knowledge graph by given its characteristics: *"A knowledge graph*

1. *Mainly describes real world entities and their interrelations, organized in a graph*
2. *Defines possible classes and relations of entities in a schema*
3. *Allows for potentially interrelating arbitrary entities with each other*
4. *Covers various topical domains."*

Another study [6] titled "Towards a Definition of Knowledge Graphs" conducted a study on the term knowledge graph and define Knowledge Graph as:

> *"A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge."*

Knowledge graphs are often differentiated based on their architecture, operational purposes, data sources, coverage and the technologies used in building them. Knowledge graphs are a key driving force for the future of artificial intelligence systems and a lot of other applications that consume and reason with structured data including search engines, enterprise and business systems, recommender systems etc.

2 Related Work

Building a Knowledge Graph is a very difficult task due to the heterogeneity of the data sources on the internet, volume or size of the data and veracity or noise in the data [1]. Knowledge graphs or knowledge base systems have been in used for some period of time. In [8], the authors show that the theory and practice of knowledge graph date back to 1982. The recent years has witnessed the evolvement of several Knowledge graphs including: Wikidata [9], YAGO [3], Freebase [13], NELL [2] PROSPERA [10] Knowledge Vault (KV) [11], Google Knowledge Graph [5], Microsoft Bing Satori [17] etc. These knowledge graphs can be classified based on their information source, scope and operational purpose. In the case of information source for example, some of the knowledge graph systems surf the internet to extract information from unstructured data sources, example of such systems include KV, NELL and PROSPERA. Other knowledge graph system may rely on human annotation and structured sources such as Freebase, or may combine the two scenarios e.g. YAGO2 [12]. In the case of scope or coverage, some focused on gathering information about a specific domain (domain specific knowledge graphs) examples include [1, 11, 15]. While others gather every information or facts across wide domains (Domain independent knowledge graphs) examples are [5, 13, 14, 17]. In the case of purpose, some of the knowledge graphs were built to be used independently such as [3, 14], while others were used as part of other systems to enhance their productivity and efficiency as it is in the case of Google Knowledge Graph and Microsoft Bing Satori.

Knowledge graphs have been built and used in other research and projects. For example in [1], a generic approach for building domain-specific knowledge graphs was proposed and this approach was employed to build a knowledge graph to combat human trafficking.

Another study [18], which complements the traditional approach of building knowledge graphs like Google's Knowledge graph focused on building event centric knowledge graph. They try to capture the dynamic state of the world by extracting information about events reported in news using state-of-the-art natural language processing and semantic web techniques. Their study also provides a method and tools to automatically build knowledge graphs from news article.

While our approach may intersect with previous methods based on information source, scope and purpose, the previous methods did not use refinement methods that improve both coverage and accuracy of knowledge graph. In addition, our proposed method will compute the correctness score for every relationship in the graph and based on that, the system can determine whether to store the newly generated knowledge after judiciously setting an accuracy threshold.

3 Proposed Architecture of the Knowledge Graph System

The architecture of the proposed system for building the knowledge graph is as shown in Fig. 2. The stages for building the knowledge graph are briefly explained below.

Data Extraction Module: This sub-system is responsible for gathering information from different sources available in open data portals through the underlying platforms application programing interfaces.

Data Analysis Module: in this stage, the information is interpreted using NLP techniques. Specifically, attempts are made to discover entities of interest from the open datasets.

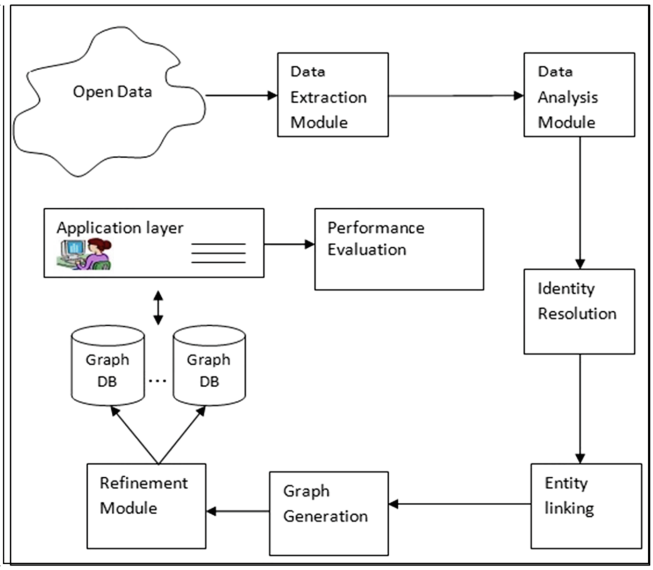

Fig. 2. Proposed architecture for the knowledge graph

Identity/Entity Resolution: in this section, we employ entity resolution methods such as Silk Link Discovery Framework [16] to resolve common entities.
Refinement Module: In this module, we improve on the quality as well as the coverage of the knowledge graph.
Performance Evaluation: This module evaluates the overall performance of the system based on some well-known gold standard graph evaluation resource.

4 Conclusions

In this work, we have considered the problem of building knowledge graph using open data. Our research agenda has the potentials to open up access to open data that are currently only accessible to a very few technical users of open data portals. Opening up access to open data as knowledge graphs will make contents of open datasets searchable using keywords or natural language phrases on existing search engines like Google. So far, only large multinational search engine providers such as Google and Microsoft provide knowledge graphs (on entities that are core to their interests) to support more intelligent search on the web. In addition our work will also significantly impact the continuous efforts of the W3C in publishing more Linked Open Data (semantically rich, open and machine readable data) on the web. Our knowledge graph approach will exploit the state of the art approach with focus on accuracy of graph relations, reasoning to discover more relations and seeking ways to increase the confidence score of relationship in the knowledge graph over time.

References

1. Szekely, P., et al.: Building and using a knowledge graph to combat human trafficking. In: Arenas, M., et al. (eds.) ISWC 2015. LNCS, vol. 9367, pp. 205–221. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25010-6_12
2. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, Jr. E.R., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: AAAI 2010, vol. 5, p. 3, July 11 2010
3. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web 2007, pp. 697–706. ACM, 8 May 2007
4. Paulheim, H.: Knowledge graph refinement: a survey of approaches and evaluation methods. In: Semantic Web Preprint, pp. 1–20 (2016)
5. Singhal, A.: Introducing the knowledge graph: things, not strings. Official Google Blog (2012)
6. Ehrlinger, L., Wöß, W.: Towards a definition of knowledge graphs. In: SEMANTiCS (Posters, Demos, SuCCESS) (2016)
7. http://opendefinition.org/. Accessed 15 Jan 2017
8. Nurdiati, S., Hoede, C.: 25 years development of knowledge graph theory: the results and the challenge (2008)
9. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)

10. Nakashole, N., Theobald, M., Weikum, G.: Scalable knowledge harvesting with high precision and high recall. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM (2011)
11. Dong, X., et al.: Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2014)
12. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. Artif. Intell. J. (2012)
13. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: SIGMOD, pp. 1247–1250. ACM (2008)
14. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
15. Schultz, A., et al.: LDIF-linked data integration framework. In: Proceedings of the Second International Conference on Consuming Linked Data, vol. 782. CEUR-WS.org (2011)
16. Isele, R., Jentzsch, A., Bizer, B.: Silk server – adding missing links while consuming linked data. In: 1st International Workshop on Consuming Linked Data (COLD 2010), Shanghai, November 2010
17. Qian, R.: Understand Your World with Bing, 21 March 2013. http://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/. Accessed 15 Jan 2017
18. Rospocher, M., et al.: Building event-centric knowledge graphs from news. Web Semant.: Sci. Serv. Agents World Wide Web **37**, 132–151 (2016)