International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217546

Using Wikipedia's Big Data for creation of Knowledge Bases

Mohamed Minhaj
Associate Professor, SDM Institute for Management Development (SDMIMD), Mysore, India

Article Info
Volume 7, Issue 6
Page Number: 11-18

Publication Issue :
November-December-2021

Article History
Accepted : 01 Nov 2021
Published : 05 Nov 2021

ABSTRACT

Wikipedia is among the most prominent and comprehensive sources of information available on the WWW. However, its unstructured form impedes direct interpretation by machines. Knowledge Base (KB) creation is a line of research that enables interpretation of Wikipedia's concealed knowledge by machines. In light of the efficacy of KBs for the storage and efficient retrieval of semantic information required for powering several IT applications such Question-Answering System, many large-scale knowledge bases have been developed. These KBs have employed different approaches for data curation and storage. The retrieval mechanism facilitated by these KBs is also different. Further, they differ in their depth and breadth of knowledge. This paper endeavours to explicate the process of KB creation using Wikipedia and compare the prominent KBs developed using the big data of Wikipedia.

Keywords: Wikipedia, WWW, IT applications

I. INTRODUCTION

The World Wide Web (WWW) is teeming with myriad forms of data. However, web users are unable to harvest and harness the vast knowledge concealed on the web because of the volume and variety of the data. Specifically, the unstructured nature of the web makes human interpretation a very challenging task. On the other hand, if the machines are used to interpret the data, it is essential that the data's context and meaning or semantics are also present.

The notion of the Semantic Web, proposed by Tim Bernes Lee, requires storing the data on the web in a form that is not only comprehensible by humans but also machine-interpretable [1]. Towards realising the Semantic Web, Knowledge Bases have a pivotal role to play. A Knowledge Base (KB) refers to an information repository that not only represents the documents and relationships between the documents but also represents data entities (ex. Person) and relationships between the entities (ex. ResidentOf). Such information repositories transform the current web of documents into a web of data.

Among the plethora of information sources available on the WWW, Wikipedia is a prominent one. Wikipedia has been the most successful collaborative encyclopaedia, and its English edition alone has around 6 million articles. However, the limited search mechanism provided on Wikipedia and the form in which its data is stored brings in many limitations for

its use by humans and direct interpretation by machines. To effectively use the knowledge concealed in Wikipedia, several attempts have been made to extract and transform the unstructured and semi-structured data of Wikipedia into structured and semantically enriched KBs. These KBs are not only enabling effective use of Wikipedia's concealed knowledge by humans, but they also facilitate better and faster interpretation of knowledge by machines. Furthermore, this has resulted in the development of many smart applications, such as Question-Answering Systems.

II. Research Objectives and Methodology

The efficacy of Knowledge Bases for storage and efficient retrieval of semantic information required for powering many information technology applications has given impetus to research on knowledge bases. Several large-scale knowledge bases, such as DBPedia, YAGO and Wikidata, have been developed. These Knowledge Bases use different data sources and are available for public or private purposes, and they employ different approaches for data curation and storage. Further, they differ in their depth and breadth of knowledge and use different mechanisms for retrieval. Besides employing existing KBs for various purposes, researchers and developers have been devising new KBs to solve specific problems.

This study endeavours to explore the domain of KBs and have the following specific objectives:

- Explicate the process of Knowledge Base creation using Wikipedia.
- Compare the prominent Knowledge Bases developed using the big data of Wikipedia.
- Identify the challenges and opportunities pertaining to the use of Wikipedia for KB generation.

As far as the study of KB creation is concerned, an exploratory method has been used. For comparison of KBs, three knowledge bases selected for the study are—Dbpedia, YAGO, and Wikidata. The selection of these Knowledge Bases was based on the criteria of being—open, large-scale, general-purpose, and Wikipedia as the primary source of knowledge.

The remainder of this paper is organised as follows: Section 3 highlights the previous essential works related to the research topic. Section 4, in the light of the scholarly literature, explains the process involved in KB creation. Section 5 presents the significance of Wikipedia for the construction of Knowledge Bases. Section 6 offers the comparison of the DBPedia, YAGO, and WikiData. Finally, section 7 presents the findings and concludes the paper.

III.Literature Review

The Knowledge Bases have a long history dating to the expert systems of the 1970s [2]. However, they have gained prominence in recent times because of their use in many semantic web applications and machine learning activities. The data in the KB is generally gathered from multiple sources and by numerous people. The collected data is integrated and stored in a structured form for easy access by users. A crucial feature of the contemporary KBs is that they are machine-interpretable; besides being used by humans, they are machine-friendly and can be employed for automated tasks. It is the computer understandability of the modern KBs, making them fit in the machine learning setting. Several KBs have been constructed. These include NELL [3], Freebase[4], OpenCYC [5], DBPedia[6], YAGO [7] and Wikidata [8].. With the difference in the knowledge sources and methods used for construction, the KBs differ in their breadth, depth, and quality of knowledge [9]. Therefore besides studying individual KBs, few research attempts have been made to compare the prominent KBs.

To compare the KBs from the perspective of Information Retrieval, Pillai et al. have empirically explored DBPedia, Wikidata, and YAGO. The study involved querying the fundamental categories of named entities like people, organisation, and location to analyse the KBs based on accessibility, timeliness, and completeness of the relations in the KBs[10].

Irrespective of the methods employed for constructing it, the resulting KB can never be perfect[9]. As the large-scale knowledge bases try to make a good trade-off between completeness and correctness, various refinement methods add missing knowledge and identify erroneous information. The work by Philipp Cimiano has compared the different approaches to Knowledge Graph Refinement[11]. This paper presents other refinement methods along with diverse evaluation methodologies. One significant finding of this work is that there are rarely any approaches that simultaneously improve the completeness and correctness of knowledge.

While large open knowledge bases or graphs contain important information, selecting one knowledge base for a particular use case is not straightforward. Depending on the task at hand and domain, some KBs may be well-fitting than others. The work by Ringler, Daniel, and Heiko Paulheim has compared the popular large-scale KBs and has attempted to provide guidelines on choosing a knowledge base that fits a given problem[12]. Their category-specific analysis has suggested different Knowledge Bases for different data requirements, like Wikidata for people, YAGO for organisations, and DBPedia for places.

IV. Building Semantic Knowledge Bases

The notion of Knowledge Bases became prominent with Expert Systems that proliferated in the 1980s due to AI research [13]. An expert system is a computer system that emulates the decision-making ability of a human expert. Expert systems are designed to solve complex problems by reasoning through bodies of knowledge, represented mainly as if-then rules rather than through conventional procedural code[14]. The two components of expert systems are the Inference Engine and the Knowledge Base. While the Knowledge Base represents facts and rules, the Inference Engine applies the rules to the known facts to deduce new facts.

The recent attention by researchers on Knowledge Base creation is because of the requirement of vast background knowledge that many contemporary information systems have for performing intelligent tasks such as Question-Answering, Natural Language Processing (NLP), and so forth. The information available today is more than at any other time in human history. However, the users cannot effectively harvest and harness the vast amount of knowledge concealed in the unstructured text. While humans can easily perceive unstructured text, it is significantly more challenging for machines to understand. This limitation has triggered the need for systems to extract information from text data automatically. Information Extraction(IE) is the task of automatically extracting information or facts from unstructured or semi-structured documents[15]. IE usually serves as a starting point for other text mining algorithms[16]. The typical sub-tasks of information extraction include - named entity recognition, coreference resolution, and relationship extraction. The named entity recognition is finding references to entities in natural language text and labelling them with their location and type[17]. While coreference resolution is the task of finding all expressions that refer to the same entity in a text[18], relation extraction involves the identification of relations between entities, such as PERSON works for

ORGANISATION (extracted from the sentence "X works for Microsoft."). IE has progressed from identifying type-of relations using manually curated text patterns [19] to domain-independent automatic discovery of relations using machine learning techniques[20]. As a result, many efficient IE systems have been developed. Whereas these systems address the very challenging and crucial task of extracting information snippets or facts from the text, the profusion of facts calls for a mechanism that can refine and organise them in a machine-friendly form. The use of ontologies is one of the popular methods to represent knowledge in a structured form suitable for machine interpretation.

Ontology is defined as a formal and explicit specification of a shared conceptualisation[21]. Conceptualisation refers to an abstract representation of the domain that we would like to model for a specific purpose. The ontology primarily contains the description of concepts, properties, and relationships among the concepts. Ontology serves as a data model to represent knowledge in the form of classes and relationships. Classes describe concepts in the domain. For example, a class of Food represents all types of Food. A class can have subclasses that represent concepts that are more specific than the superclass. For instance, we can divide the class of Food into Fruit and Vegetable. Fruit, in turn, can have many subclasses such as Citrus and Fleshy. Specific Food such as Orange can be considered as an instance of the class - Citrus.

Relations in an ontology specify how objects are related to other objects. Typically a relationship or relation is of a particular type/class that conveys in what way an object is related to another object in the ontology. For example, the ontology that contains the concept "Apple" and the concept "Red" might be related by a relation of type "HasColour."

An ontology basically provides a common vocabulary for users who need to share information in a domain. Information extracted from textual sources and stored in ontologies serves as KBs that enable computers to understand the semantics and simulate intelligent behaviour [21].

V. Wikipedia, the Big Data for Knowledge Base Creation

The content in Wikipedia is available in 300 languages, and the English edition alone has around 6 million articles. Wikipedia is based on a model of openly editable content. Wikipedia's big data of text, facts, and figures are open for all its users to edit and contribute to the fast-growing online repository. The Wikipedia contributors, in addition to the quantity, strive to improve the quality. Every article in Wikipedia is considered to be in a work-in-progress phase and progresses to various stages of completion. As articles develop, they tend to become more comprehensive and balanced. Quality also improves over time as misinformation and other errors are removed or rectified.

In addition to a large amount of unstructured data, Wikipedia articles also have structured data enveloped in them in the form of Infoboxes (Depicted in Figure 2). An Infobox is a fixed-format table usually added to the top right-hand corner of articles. The infoboxes summarise essential points in an easy-to-read format and also improve navigation to other related articles. With the structured nature of data and the possibility of mapping its schema efficiently to many dominant metadata systems, the data in Wikipedia's infoboxes is widely used for many knowledge-based applications. The notable applications that are built primarily on Wikipedia's infoboxes include DBPedia.

Figure 1 Infobox in Wikipedia

VI. The comparison of Prominent KBs built using Wikipedia

In light of the wide applications of Knowledge Bases in NLP and AI tasks, several Knowledge Bases have been constructed. A comparison of three prominent large-scale, open, and general-purpose Knowledge Bases with Wikipedia as the primary source of knowledge is presented below. Table 1 gives key information about the three knowledge bases factored in the present study.

DBPedia

The primary objective of the DBpedia project was to convert Wikipedia content into structured knowledge, which could facilitate the use of Semantic Web techniques, such as asking sophisticated queries against Wikipedia, linking it to other datasets on the web, or creating new applications. The project developed an information extraction framework to convert Wikipedia content into RDF and consisted of 103 million RDF triples[6]. Besides developing a web interface to access the knowledge base, the open and linked nature of the DBpedia facilitates linking its content to other open KBs and integration with other semantic technologies. After its initial version was developed in 2007, a large global community has continuously improved and extended the DBpedia project. DBpedia has been the precursor of many successful KBs that are used today. Many projects have employed DBpedia for prototyping and proofs-of-concept and have been instrumental in many semantic technology innovations. Many enterprises, such as Apple, Google, and IBM, have adopted the idea of data extraction from DBPedia for their high-visibility AI projects, Siri, Google Knowledge Graph, and Watson, respectively.

Wikidata

Wikimedia Foundation started the Wikidata project to create a central storage for the structured data of its sister projects, including Wikipedia. With the

collaborative and multilingual nature of Wikipedia, the same piece of information appears in articles in many languages. Further, the same piece of information also appears in many articles within a single language collection. For example, besides being available in both English and Italian articles about Rome, Rome's population is also available in the English article Cities in Italy. However, the numbers on all these pages are different. The goal of Wikidata is to overcome these problems by creating new ways for Wikipedia to manage its data on a global scale. Wikidata, launched in 2012, is one of the widely used, free, and open knowledge bases that can be read and edited by both humans and machines.

YAGO

The motivation behind the YAGO project was the development of a huge ontology with knowledge from several sources instead of relying on a single source of background knowledge. While the core of YAGO was assembled from Wikipedia, but rather than using information extraction methods to leverage the knowledge of Wikipedia, YAGO utilises the category pages of Wikipedia. Category pages with lists of articles that belong to a specific category were used to get the candidates for entities, concepts, and relations. One of the critical requirements of any ontology is the arrangement of concepts in a taxonomy. Though Wikipedia categories are arranged in a hierarchy, they are not very useful for ontological purposes. For example, in Wikipedia, the information about a football player X, who is a citizen of country C, may have a super-category named "Football in C", wherein X may be interpreted as a Football and not as a player. WordNet, in contrast, provides a clean and carefully assembled hierarchy of thousands of concepts. However, Wikipedia concepts have no apparent counterparts in WordNet. YAGO project developed a technique that links two sources with near-perfect accuracy. This approach enabled YAGO to use the vast number of individuals known in Wikipedia, coupled with the exploitation of clean taxonomy of concepts from WordNet. Since its initial Release in 2008, YAGO has undergone several improvements, and because of the ontology's high coverage and high quality, it has been employed for several AI systems. A prominent application of YAGO is its use in the IBM Watson.

	DBpedia	**YAGO**	**Wikidata**
Developed by	Leipzig University University of Mannheim OpenLink Software	Max-Planck-Institute, Saarbrücken	Wikimedia Foundation
Website	https://www.dbpedia.org/	https://yago-knowledge.org/	https://www.wikidata.org/
Year of Launch	2007	2008	2012
License	Creative Commons Attribution-ShareAlike 3.0 License and the GNU Free Documentation License.	Creative Commons Attribution 4.0 International License	Creative Commons CC0 License and Creative Commons Attribution-ShareAlike License;
Sources of data	Wikipedia	Wikipedia, WordNet, GeoNames and	Wikipedia

		Schema.org	
Size of the KB as on April 2021 (No. of Entities /Things/Items)	Around 38 million (Source : https://wiki.dbpedia.org/about)	Around 50 million (Source : https://yago-knowledge.org/getting-started#what-is-yago)	Around 93 million (Source : https://www.wikidata.org/wiki/Wikidata:Main_Page)

Table 1. Key information about DBPedia, YAGO and Wikidata

VII. CONCLUSION

Knowledge Bases serve as a backbone for many contemporary IT applications. A knowledge base is a machine-comprehensible collection of knowledge about the real world. It fundamentally contains entities (such as people and organisations) and their relations (such as birthPlace, director, etc.). Among the plethora of sources available on WWW for developing Knowledge Bases, Wikipedia is widely used because of its open and collaborative nature in assimilating the world's knowledge. Several research projects have leveraged Wikipedia for building Knowledge Bases.

Further, these knowledge bases have been employed for many successful industrial applications, including IBM's Jeopardy-winning Watson system. This exploratory study found that all the prominent KBs such as DBPedia, YAGO and WIKIDATA have a great potential for semantic applications. However, as they differ in their depth and breadth of knowledge, choosing a particular knowledge base for a given problem is a non-trivial task. Further, when an organisation's specific requirements are not satisfied by the existing KBs, organisations can either enhance the existing KBs or create new KB by leveraging the big data of Wikipedia.

VIII. REFERENCES

[1]. T. Berners-Lee, J. Hendler, and O. Lassila, 'The semantic web', Sci. Am., vol. 284, no. 5, pp. 34–43, 2001.

[2]. A. Ratner and C. Ré, 'Knowledge Base Construction in the Machine-learning Era: Three critical design points: Joint-learning, weak supervision, and new representations', Queue, vol. 16, no. 3, pp. 79–90, Jun. 2018, doi: 10.1145/3236386.3243045.

[3]. A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka, and T. Mitchell, 'Toward an architecture for never-ending language learning', in Proceedings of the AAAI Conference on Artificial Intelligence, 2010, vol. 24, no. 1.

[4]. K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, 'Freebase: a collaboratively created graph database for structuring human knowledge', in Proceedings of the 2008 ACM SIGMOD international conference on Management of data, Vancouver, Canada, Jun. 2008, pp. 1247–1250. doi: 10.1145/1376616.1376746.

[5]. D. B. Lenat, 'CYC: a large-scale investment in knowledge infrastructure', Commun. ACM, vol. 38, no. 11, pp. 33–38, Nov. 1995, doi: 10.1145/219717.219745.

[6]. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, 'Dbpedia: A nucleus for a web of open data', in The semantic web, Springer, 2007, pp. 722–735.

[7]. F. M. Suchanek, G. Kasneci, and G. Weikum, 'Yago: a core of semantic knowledge', in

Proceedings of the 16th international conference on World Wide Web, 2007, pp. 697–706.

[8]. D. Vrandečić and M. Krötzsch, 'Wikidata: a free collaborative knowledgebase', Commun. ACM, vol. 57, no. 10, pp. 78–85, 2014.

[9]. A. Bordes and E. Gabrilovich, 'Constructing and mining web-scale knowledge graphs: KDD 2014 tutorial', in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, New York, USA, Aug. 2014, p. 1967. doi: 10.1145/2623330.2630803.

[10]. S. G. Pillai, L.-K. Soon, and S.-C. Haw, 'Comparing DBpedia, Wikidata, and YAGO for web information retrieval', in Intelligent and Interactive Computing, Springer, 2019, pp. 525–535.

[11]. H. Paulheim, 'Knowledge graph refinement: A survey of approaches and evaluation methods', Semantic Web, vol. 8, no. 3, pp. 489–508, 2017.

[12]. D. Ringler and H. Paulheim, 'One knowledge graph to rule them all? Analyzing the differences between DBpedia, YAGO, Wikidata & co.', in Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz), 2017, pp. 366–372.

[13]. C. T. Leondes, Ed., in Expert Systems: The Technology of Knowledge Management and Decision Making for the 21st Century, 1st edition., San Diego: Academic Press, 2001, pp. 1–22.

[14]. P. Jackson, in Introduction to Expert Systems, 3rd ed., USA: Addison-Wesley Longman Publishing Co., Inc., 1998, p. 2.

[15]. J. Cowie and W. Lehnert, 'Information Extraction', Commun ACM, vol. 39, no. 1, pp. 80–91, Jan. 1996, doi: 10.1145/234173.234209.

[16]. M. Allahyari et al., 'A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques', ArXiv170702919 Cs, Jul. 2017, Accessed: Sep. 26, 2018. Online]. Available: http://arxiv.org/abs/1707.02919

[17]. R. Leaman and G. Gonzalez, 'BANNER: an executable survey of advances in biomedical named entity recognition', in Biocomputing 2008, World Scientific, 2008, pp. 652–663.

[18]. The Stanford Natural Language Processing Group'. https://nlp.stanford.edu/projects/coref.shtml (accessed Jun. 20, 2020).

[19]. M. A. Hearst, 'Automatic acquisition of hyponyms from large text corpora', presented at the The 15th international conference on computational linguistics, 1992.

[20]. M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, 'Open information extraction from the web.', in Proceedings of the International Joint Conference on Artificial Intelligence, 2007, pp. 2670–2676.

[21]. T. R. Gruber, 'Toward principles for the design of ontologies used for knowledge sharing?', Int. J. Hum.-Comput. Stud., vol. 43, no. 5–6, pp. 907–928, 1995.

Cite this article as :

Mohamed Minhaj, "Using Wikipedia's Big Data for creation of Knowledge Bases", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 6, pp. 11-18, November-December 2021. Available at doi : https://doi.org/10.32628/CSEIT217546
Journal URL : https://ijsrcseit.com/CSEIT217546