Royal Netherlands Academy of Arts and Sciences (KNAW)

KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN

Unflattening Knowledge Graphs

Van Erp, Marieke

2023

DOI (link to publisher)
10.1145/3587259.3630082

document version
Publisher's PDF, also known as Version of record

Link to publication in KNAW Research Portal

citation for published version (APA)
Van Erp, M. (2023). *Unflattening Knowledge Graphs*. 223-224. https://doi.org/10.1145/3587259.3630082

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the KNAW public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the KNAW public portal.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
pure@knaw.nl

Download date: 12. Sep. 2024

Unflattening Knowledge Graphs

Marieke van Erp
marieke.van.erp@dh.huc.knaw.nl
DHLab, KNAW Humanities Cluster
Amsterdam, Netherlands

ABSTRACT

Large general-purpose knowledge graphs (KGs) are a critical component for knowledge-driven applications. However, most KGs represent only a limited view of the entities and concepts they describe. The concept coffee can, for example, refer to the plant that yields coffee seeds, the beverage 'coffee', and the activity of drinking the beverage. Moreover, it has a long history that is deeply connected to colonialism and status. All of these notions are an intricate part of national identities, have changed dramatically over time, and connect to many different narratives with different opinions on them. This complexity is not captured in current KGs. In this vision paper, I present the three crucial challenges for unflattening knowledge graphs and directions for future work.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; **Semantic networks**.

KEYWORDS

digital humanities, knowledge graphs, language technology

ACM Reference Format:
Marieke van Erp. 2023. Unflattening Knowledge Graphs. In *K-CAP 2023: The Twelfth International Conference on Knowledge Capture, December 05-07, 2023, Pensacola, Fl, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/https://doi.org/10.1145/3587259.3630082

1 INTRODUCTION

Large knowledge graphs (KGs) such as Wikidata and DBpedia provide a wealth of information that is used by many applications. Whilst they have undeniably impacted knowledge-intense data-driven systems, they only express a limited representation of the concepts and entities they represent [15]. To address a wider range of applications especially in the humanities and social science, we need to deepen concept and entity representations to capture more of their complexity in knowledge graphs.

For example, at the time of writing, Wikidata represents coffee as 'brewed beverage made from seeds of Coffea genus' and as a subclass of 'drink', 'stimulant foodstuff', 'coffee drink', 'soft drink', 'hot beverage', 'non-alcoholic beverage' and 'colonial goods'. Only that last category hints at information regarding this concept that is not related to its qualities as a beverage. DBpedia similarly focuses its description on the food-dimension. A popular application of KGs is to use them as targets for entity linking. When automatically linking the a reference in a text to 'coffee' to a KG, the concept is currently underspecified so we cannot easily discern references to the drink, the plant that yields the coffee seeds or the activity of drinking it, which can have different meanings in different social contexts. Moreover, references to its colonial history as for example described in the text of the Wikipedia page on coffee are currently not yet available in the structured Wikidata representation.

Being able to automatically distinguish between these different dimensions of coffee, or for that matter and multi-dimensional concept, is useful for entity linking or many other tasks like search or visualization. To be able to do so, we need to **unflatten KGs to include more multidimensional aspects of entities**. Specially, there are three critical challenges that need to be addressed in order to perform this unflattening: development of better identity representations; better approaches for capturing change over time as concepts and entities evolve; and dealing with the long tail of entities and concepts. In the remainder of this paper, I will discuss each of these challenges in turn.

2 CHALLENGE 1: IDENTITY

The semantic web research community is aware that knowledge graphs currently lack a certain precision in dealing with identity and change over time (cf. [3]). A fair body of research focuses on identity and logical equality on semantic web datasets to assess whether one thing which has two (or more) names should be considered the same entity or concept [4]. Various analyses of the use of owl:sameAs, the main relationship used to express that two entities or concepts are the same, have shown that in many cases the identity criterion of logical equality is not abided by, thus generating factoids that equate for example a general description of the Netherlands in 2020 as found on Wikipedia, to a description of the Dutch Republic in 1750 in a historical database. In certain contexts, these entities can be treated as being the same, but not in all contexts [4, 12].

In the semantic web, the way owl:sameAs seems to be applied in certain cases, seems to correspond with the linguistic concepts coreference and near-identity [13], where two names are considered logically identical within a given context. In these cases, the creators had a subset of properties of the resources in mind on the basis of which the identity link was established, for example during a particular time, the ruler of Spain was Franco, but only in the context 1939 - 1975. To resolve some of these issues [8], have proposed an identity ontology. [1] further build on this and compute identity relationships over sets of properties instead of all properties.

Current language technology methods such as [19] achieve high performance in extracting entities and relationships. To create richer identity profiles for concepts and entities that can begin

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
K-CAP 2023, December 05-07, 2023, Pensacola, Fl, USA
© 2023 Association for Computing Machinery.
ACM ISBN 979-8-4007-0141-2/23/12…$15.00
https://doi.org/https://doi.org/10.1145/3587259.3630082

to capture some of the contextual information that discerns their use in different settings, we need to step up our use of these technologies. Preliminary work in [15] shows that parts of KGs can be aid in this. Next, linkers need to be made more semantically-aware, and for example filter out nonsensical answers, as proposed in [20].

3 CHALLENGE 2: CHANGE

Concepts and entities change over time: Arnold Schwarzenegger has been a bodybuilder, actor, politician, director, restaurateur, writer, and soldier. For some entities and concepts, this information is encoded in KGs, but the formats and the extent to which the different properties are encoded and during what timeframe vary.

The phenomenon of changing meaning of concepts is known as concept drift in semantic web and computational linguistics research [14]. Detection of concept drift in texts using distributional semantic models have shown that it is possible to track changes in vocabulary over time (cf. [7]). Limitations here are that the approach depends on the availability of large amounts of manually labelled data. Unsupervised clustering methods [10] and concept networks [9] provide less data-intensive approaches.

An entity or concept's label, intension, and extension [18] define its identity (including its entity space and contexts). As knowledge engineers, we need to identify and model explicitly how an identity evolves, including where it splits, merges or ceases to exist and relates to other entities and concepts. Without this, our KGs remain fixed snapshots describing entities frozen in time.

4 CHALLENGE 3: THE LONG TAIL

In many technological advances, the Pareto principle applies: roughly 80% of the effects are covered by 20% of the causes. In entity recognition and linking, this translates to a small set of entities being mentioned often (the 'head' entities) and a large portion of entities being mentioned (relatively) infrequently (the 'long tail' entities) [17]. Because evaluation datasets are biased towards head entities focusing on the head of the distribution generally yields high performances [6, 16].

To rediscover properties of long tail entities and concepts and fill in the gaps where not enough information is available, open information extraction from different sources is key. Knowledge enrichment, as presented in [2] can leverage 'known' information from similar entities and concepts to guide property extraction in the long tail. We need to explicitly model and store the provenance of the information source, such that less well-represented entities and their properties can be retrieved instead of them disappearing into the vastness of the extracted and inferred data.

5 WHERE TO GO FROM HERE

Taking on the issues of identity, change, and the long tail is not trivial. Fortunately, many pieces of the puzzle are already in place: to a certain extent the community is aware of the gaps, but more attention is needed. Language technology has advanced to a stage where it can be employed to detect fine-grained and contextual entities and relationships, but we need to pair them with precise and wide-ranging evaluation datasets to further the usability of them in multiple settings. In particular in combination with already available structured data high quality broad coverage KG improvements can be enacted. Now is the time to develop methods to unflatten knowledge graphs.

ACKNOWLEDGMENTS

Funded by the European Union under grant agreement 101088548 - TRIFECTA. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

[1] Wouter Beek, Stefan Schlobach, and Frank van Harmelen. 2016. A contextualised semantics for owl: sameas. In *The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29–June 2, 2016, Proceedings 13*. Springer, 405–419.

[2] Ermei Cao, Difeng Wang, Jiacheng Huang, and Wei Hu. 2020. Open knowledge enrichment for long-tail entities. In *Proceedings of The Web Conference 2020*.

[3] Melisachew Chekol, Giuseppe Pirrò, Joerg Schoenfisch, and Heiner Stuckenschmidt. 2017. Marrying uncertainty and time in knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[4] Harry Halpin, Patrick J Hayes, James P McCusker, Deborah L McGuinness, and Henry S Thompson. 2010. When owl: sameas isn't the same: An analysis of identity in linked data. In *The Semantic Web–ISWC 2010: 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I 9*. Springer, 305–320.

[6] Filip Ilievski, Piek Vossen, and Stefan Schlobach. 2018. Systematic study of long tail phenomena in entity linking. In *Proceedings of the 27th international conference on computational linguistics*. 664–674.

[7] Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten De Rijke. 2015. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM international on conference on information and knowledge management*.

[8] Jamie P McCusker and Deborah L McGuinness. 2010. Towards Identity in Linked Data.. In *OWLED*.

[9] Adina Nerghes, Ju-Sung Lee, Peter Groenewegen, and Iina Hellsten. 2014. The shifting discourse of the European Central Bank: Exploring structural space in semantic networks. In *2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems*. IEEE, 447–455.

[10] Nidhi, Veenu Mangat, Vishal Gupta, and Renu Vig. 2018. Methods to investigate concept drift in big data streams. *Knowledge Computing and Its Applications: Knowledge Manipulation and Processing Techniques: Volume 1* (2018), 51–74.

[12] Joe Raad, Wouter Beek, Frank Van Harmelen, Nathalie Pernelle, and Fatiha Saïs. 2018. Detecting erroneous identity links on the web using network metrics. In *The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I 17*. Springer, 391–407.

[13] Marta Recasens, Eduard Hovy, and M Antònia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua* 121, 6 (2011).

[14] Martin Scholz and Ralf Klinkenberg. 2007. Boosting classifiers for drifting concepts. *Intelligent Data Analysis* 11, 1 (2007), 3–28.

[15] Marieke van Erp and Paul Groth. 2020. Towards Entity Spaces. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France.

[16] Marieke Van Erp, Pablo Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Jörg Waitelonis. 2016. Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 4373–4379.

[17] Andreas Vlachidis and Douglas Tudhope. 2016. A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain. *Journal of the association for information science and technology* 67, 5 (2016).

[18] Shenghui Wang, Stefan Schlobach, and Michel Klein. 2011. Concept drift and how to identify it. *Journal of Web Semantics* 9, 3 (2011), 247–265.

[19] Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Automated Concatenation of Embeddings for Structured Prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2643–2660.

[20] Konstantin Ziegler, Olivier Caelen, Mathieu Garchery, Michael Granitzer, Liyun He-Guelton, Johannes Jurgovsky, Pierre-Edouard Portier, and Stefan Zwicklbauer. 2017. Injecting semantic background knowledge into neural networks using graph embeddings. In *2017 IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*. IEEE, 200–205.