Towards a Better Contextualization of Web Contents via Entity-Level Analytics

Amit Kumar(✉)

Department of Computer Science, Université de Caen Normandie,
Campus Côte de Nacre, 14032 Caen Cedex, France
amit.kumar@unicaen.fr

Abstract. With the abundance of data and wide access to the internet, a user can be overwhelmed with information. For an average Web user, it is very difficult to identify which information is relevant or irrelevant. Hence, in the era of continuously enhancing Web, organization and interpretation of Web contents are very important in order to easily access the relevant information. Many recent advancements in the area of Web content management such as classification of Web contents, information diffusion, credibility of information, etc. have been explored based on text and semantic of the document. In this paper, we propose a purely semantic contextualization of Web contents. We hypothesize that named entities and their types present in a Web document convey substantial semantic information. By extraction of this information, we aim to study the reasoning and explanation behind the Web contents or patterns. Furthermore, we also plan to exploit LOD (Linked Open Data) to get a deeper insight of Web contents.

Keywords: Contextualization · Knowledge extraction · Entity-level analytics

1 Introduction and Motivation

Even in the 30th year of World Wide Web, we can still observe the enormous amount of growth in the Web contents being created and subsequently available to any internet user. Because of the inconsistency of the data being generated, it is very hard for an ordinary user to distinguish the Web contents according to their societal relevance. With the availability of NER techniques [5] and LOD [1,11,13], we have access to a lot of information about the named entities described in a Web content. This contextualization of the entities contained in a text can help us to deal with Web contents. We observe that, for a text describing an event, there are specific recurring patterns of entity types appearing together. For instance, in the case of ‘natural disasters’, entities like organizations, countries, presidents appear together whereas in the case of ‘political events’, entities like parties, leaders, business-persons appear together. The availability of tools

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 613–618, 2020.
https://doi.org/10.1007/978-3-030-45442-5_80

like AIDA [16] or DBPedia Spotlight [12], which can interlink text documents to LOD has provided us efficient means to capture the semantics of a plain text using the entity-level.

The purpose of this study is to analyze those large amounts of data and to help the user in getting a better semantic understanding of a Web document. To this end, we aim to study a Web document semantically using entity-level analytics. Ultimately, we plan to exploit and aggregate external knowledge using LOD for the proper contextualization of a Web content.

2 Background and Related Work

Knowledge bases (KBs) are an effective way to store Web documents semantically in a structured format. Because of easy accessibility, these KBs are fruitful resources for many tasks in information retrieval [4] and natural language processing [9]. Recently, researchers from different domains have developed different knowledge acquisition approaches for the creation of knowledge graphs. This results in an emanation of large publicly accessible KBs such as Freebase [1], DBPedia [11], YAGO [13], which accommodate spatial and temporal information in addition to structural knowledge. Many applications such as complex event detection [17], named entities disambiguation [14] or social media topic classification [3] from various domains have acquired the benefit by integrating knowledge from LOD.

Entity-level analytics aggregates semantic information by incorporating knowledge about an entity or its types. The problem of event diffusion prediction into foreign language communities [8] has shown encouraging results with the assimilation of knowledge about the entities contained in a document. Here the introduced framework ELEVATE only utilizes the information about the entities in the document and resources from YAGO [13]. In [7], the authors address the task of Web content fine-grained hierarchical classification. They hypothesize that a document is symbolized by the named entities it comprised. They propose the idea of the 'semantic fingerprinting' method that expresses the overall semantics of a Web document by a compact vector. Entity-level analytics is also effective in computational fact checking of information [2]. The authors claim that human fact checking can be achieved by finding the shortest path on a conceptually or semantically defined network such as knowledge graphs (KGs).

Entity-level analytics provide a depth insight into Web contents. KGs carry a lot of information about entities, but all the information is not equally important for a given text. The novelty of this thesis is to discriminate between interesting and uninteresting semantic information about entities w.r.t. the context of a text.

3 Current Work

3.1 CALVADOS : For Entity-Level Content Analysis

The idea of semantic fingerprinting [6] - as an approach towards Web analytics was well acknowledged by researchers in the semantic Web community. Thus, we presented the CALVADOS system [7] as an extension of semantic fingerprinting. At first, this system filters all the named entities present in a given text. By utilization of type information for all these entities from YAGO, it creates a representative vector (called semantic fingerprint) for the text. In last, it predicts the fine-grained type of the content using machine learning techniques. Moreover, it reports the semantic building block of the text. Figure 1 outlines the conceptual pipeline. The notable contributions of the mentioned scientific article are:

- employ `semantic fingerprint` to represent document's semantics
- exploration and visualization of dependencies among entities comprised
- data digestion supported by providing contextual KB data (e.g., types)

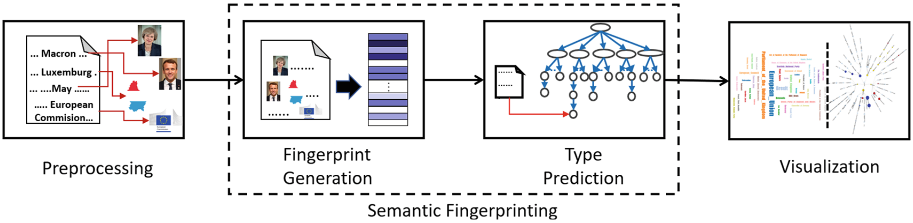

Fig. 1. Conceptual overview of the CALVADOS pipeline

3.2 Concise Entity-Type Extraction

Exploitation of named entities and their types are always valuable in getting a better contextualization of a Web document [7]. But, sometimes we can be easily drowned by too much information. For example, recent articles involving Donald Trump deal with his position of president. But Trump has 76 facets (considering Wordnet types) like communicator, president, business-person, etc., which are not equally relevant. For some entities, it is even more complicated. For instance, Arnold Schwarzenegger is famous to be an actor, a politician and a bodybuilder. When an article deals with him, the context of the article makes us understand which facet is relevant, e.g., 'actor' if the article deals with a film release. Hence, it arises two research questions:

- **RQ1:** What are the most relevant type(s) for an entity in general (i.e., without context)?
- **RQ2:** What are the most relevant type(s) for an entity in a given context?

Computational Model: Let $d_i \epsilon D$ represents a document. The named entities associated with a document d_i is given by $N(d_i)$. $T(n_j)$ represents all the k types associate with any entity n_j. Entity-level document type is represented by d_i^t as shown in Eq. 4. Our task is to select m number of types from $T(n_j)$, where

$m << k$. We define two types of models - First, for the calculation of T_{Gen} (*i.e.* **RQ1**, Eq. 5) and second, for the calculation of T_{Con} (*i.e.* **RQ2**, Eq. 6).

$$D = d_1, d_2.....d_p \tag{1}$$

$$N(d_i) = n_1^i, n_2^i, ...n_q^i \tag{2}$$

$$T(n_j) = t_1^j, t_2^j, ...t_k^j \tag{3}$$

$$\text{Entity-level typed document}, d_i^t = \varphi(n_1^i, n_2^i, ...n_q^i) \tag{4}$$

$$T_{Gen} = f_{gen}(T(n_j)) \tag{5}$$

$$T_{Con} = f_{con}(T(n_j), text) \tag{6}$$

Currently, we have focus on our first research question (**RQ1**). Our first challenge is to find or develop the appropriate data set for the aforementioned task.

Gold Standard Creation: It is not easy to find the most relevant type(s) for an entity in general. We create the gold standard based on the Wordnet hierarchy mentioned in YAGO (1981 types). We consider that the most relevant type(s) for any entity is mentioned in its Wikipedia page. Precisely, we extract the types that are mentioned in the first or second sentence of entity's Wikipedia page and map them to the mentioned hierarchy.
Example – Extracted Wikipedia labels for 'Arnold Schwarzenegger' are actor, filmmaker, businessman, author, bodybuilder and politician. After mapping of these labels to Schwarzenegger's Wordnet hierarchy in YAGO, the ground truths are *actor, film-maker, businessman, bodybuilder and politician.*

Experimental Pipeline: Our next challenge is to find the suitable mechanism for concise entity-type prediction in general. We rely only on the structural information, which we get by exploring knowledge graph of entity in YAGO. We implemented several techniques as baselines but none of these techniques show promising results. These models are:

- **Based on Leaf Node:** We had the intuition that the most specific or relevant type of an entity should be at the deepest in the YAGO Wordnet entity's hierarchy. So, we picked the type that is at the deepest in the hierarchy.
- **Based on Branching Factor:** While implementing the model based on leaf node, we observe that sometimes, we were selecting a too specific type, e.g., forward (child of football-player in the hierarchy) instead of football-player. So, we decided to pick the node that has the highest branching factor (number of direct children) and at the deepest in the hierarchy.
- **Based on ML Classifier:** We developed a model based on random forest. We used all the Wordnet types present in the entity's hierarchy as features.

We aim to develop a relevant types prediction model based on Graph Neural Network [15]. More specifically, we utilize the concept of Graph Convolutional Network (GCN) [10]. While implementation, we faced the following challenges:

- In GCN, Readout function [15] is used to get embedding for the graph based on an aggregation of node features from the final iteration. Finding the suitable Readout function for our task is one of the main challenges.
- Entity's graph is a sub-graph of YAGO Wordnet hierarchy. Only delivering the structure of the sub-graph is not sufficient. It needs label information along with the structure of sub-graph. Encoding node label is our next challenge. One hot encoding is one of the solutions for giving the label information.

4 Challenges and Next Steps

Based on our proposed research and current work progress, we find the following challenges to handle in the near future:

- In the early stage of our experiments, we realise that some of the types within a category are very hard to predict, e.g., in person category - there are entities with ground truth types 'intellectual' or 'military officer' where the model fails to predict it correctly. Our challenge is to find the common patterns among these sub-categories and to propose the solution for this failure.
- Our next challenge is to develop a gold standard for task 2 (**RQ2**).
- Our last challenge is to develop method for types prediction in a given context.

References

1. Bollacker, K., et al.: Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of SIGMOD 2008, pp. 1247–1250. ACM (2008)
2. Ciampaglia, G.L., et al.: Computational fact checking from knowledge networks. PLoS ONE **10**(6), 1–13 (2015)
3. Cano, A.E., et al.: Harnessing linked knowledge sources for topic classification in social media. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media, HT 2013, pp. 41–50. ACM (2013)
4. Dalton, J., Dietz, L., Allan, J.: Entity query feature expansion using knowledge base links. In: Proceedings of SIGIR 2014, pp. 365–374. ACM (2014)
5. Finkel, J. R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of ACL 2005, pp. 363–370. ACL (2005)
6. Govind, Alec, C., Spaniol, M.: Semantic fingerprinting: a novel method for entity-level content classification. In: Mikkonen, T., Klamma, R., Hernández, J. (eds.) ICWE 2018. LNCS. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91662-0_21
7. Govind, Kumar, A., Alec, C., Spaniol, M.: CALVADOS: a tool for the semantic analysis and digestion of web contents. In: Hitzler, P., et al. (eds.) ESWC 2019. LNCS, pp. 1–6. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32327-1_17
8. Govind, Spaniol, M.: ELEVATE: a framework for entity-level event diffusion prediction into foreign language communities. In: Proceedings of WebSci 2017, pp. 111–120. ACM (2017)

9. Hao, Y., et al.: Pattern-revising enhanced simple question answering over knowledge bases. In: Proceedings of COLING 2018, pp. 3272–3282. ACL (2018)
10. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations, ICLR 2017 (2017)
11. Lehmann, J., et al.: DBpedia-a large-scale, multilingual knowledge base extracted from Wikipedia. Semant. Web J. **6**, 167–195 (2015)
12. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of I-Semantics 2011, pp. 1–8. ACM (2011)
13. Rebele, T., Suchanek, F., Hoffart, J., Biega, J., Kuzey, E., Weikum, G.: YAGO: a multilingual knowledge base from Wikipedia, Wordnet, and Geonames. In: Groth, P., et al. (eds.) ISWC 2016. LNCS, vol. 9982, pp. 177–185. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46547-0_19
14. Usbeck, R., et al.: AGDISTIS - graph-based disambiguation of named entities using linked data. In: Mika, P., et al. (eds.) ISWC 2014. LNCS, vol. 8796, pp. 457–471. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11964-9_29
15. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? In: International Conference on Learning Representations, ICLR 2019 (2019)
16. Yosef, M.A., et al.: AIDA: an online tool for accurate disambiguation of named entities in text and tables. In: Proceedings of VLDB, vol. 2011, pp. 1450–1453 (2011)
17. Yan, Y., et al.: Event oriented dictionary learning for complex event detection. IEEE Trans. Image Process. **24**(6), 1867–1878 (2015)