A Survey on Challenges for Entity Retrieval in Web Markup Data

Ran Yu, Besnik Fetahu, Ujwal Gadiraju and Stefan Dietze

L3S Research Center, Leibniz Universität Hannover, Germany
{yu, gadiraju, fetahu, dietze}@L3S.de

Abstract. Embedded markup based on Microdata, RDFa, and Microformats have become prevalent on the Web and constitute an unprecedented data source. RDF statements from markup are highly redundant, co-references are very frequent yet explicit links are missing, and frequently contain errors.
We present a preliminary analysis on the challenges associated with markup data in the context of entity retrieval. We analyze four main factors: (i) *co-references*, (ii) redundancy, (iii) inconsistencies, and (iv) accessibility of information in the case of URLs. We conclude with general guidelines on how to handle such challenges when dealing with embedded markup data.

1 Introduction

Markup annotations embedded in HTML pages have become prevalent on the Web, building on standards such as RDFa[1], Microdata[2] and Microformats[3], and driven by initiatives such as schema.org, a joint effort led by Google, Yahoo!, Bing and Yandex.

The Web Data Commons [2], a recent initiative investigating a Web crawl of 2.01 billion HTML pages from over 15 million pay-level-domains (PLDs) found that 30% of all pages contain some form of embedded markup already, resulting in a corpus of 20.48 billion RDF quads. The scale of the data suggests potential for a range of tasks, such as entity retrieval, knowledge base population, or entity summarisation. Initial case studies investigate for instance the scope of bibliographic data [4] or learning resources metadata [5] sourced from Web markup.

However, entities described through extracted facts from embedded markup have different characteristics compared to Linked Data. For example, co-references are very frequent, yet not linked through explicit statements. In addition, statements are highly redundant and often limited to a small set of highly popular predicates, complemented by a long tail of less frequent predicates. Moreover, the extracted data contains a wide variety of syntactical and semantic errors.

In this work, we present an overview of challenges and attributes of markup data. We focus on the *entity retrieval* use case which represents one of the prerequisites for tasks like *entity summarization* [6] or *knowledge base augmentation* [3].

2 Web Markup Attributes and Challenges

In this section, we present a preliminary analysis of some of the key attributes of markup data, and the challenges that are associated with them. We focus on the use of markup data for *entity retrieval*.

[1] RDFa W3C recommendation: http://www.w3.org/TR/xhtml-rdfa-primer/
[2] http://www.w3.org/TR/microdata
[3] http://microformats.org

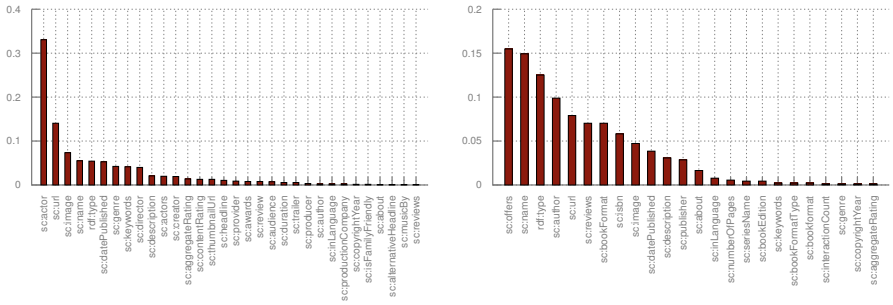
We conduct a preliminary analysis for two entity types extracted from the WDC 2014 dataset, `Movie` and `Book`, since they correspond to a considerable amount of data and are easy validate manually. For each type we randomly pick 10 entity queries which we select from Wikipedia entity names from the respective categories. We setup a Lucene[4] index for the different subsets of markup data for the two types. We retrieve top–500 entities by querying on the predicate `sc:name` with BM25 as our *query similarity* model. We manually evaluate the corresponding result sets, and find out that there are no relevant entities beyond top–50 (see Table 1 for details). In more detail, we analyze the following aspects of Web Markup data.

Entity Co-References. We investigate the occurence of co-references, namely that of resource descriptions that are about the same entity. In contrast to traditional RDF datasets, statements extracted from markup form sparsely connected graphs. In our case, none of the *relevant entities* are interlinked explicitly. From our evaluation for the types `Movie` and `Book`, we find the following number of relevant entities for our random query set.

type	*q1*	*q2*	*q3*	*q4*	*q5*	*q6*	*q7*	*q8*	*q9*	*q10*
Movie	0	2	6	4	3	2	0	5	46	0
Book	13	1	0	40	28	15	11	4	1	0

Table 1: Amount of relevant entity descriptions for each query for types `Movie` and `Book`.

Information Redundancy. Here we deal with the *redundancy* aspect of statements from Web markup. From the relevant entities for our query set, we have a set of 42 distinct predicates. The distribution of statements for each predicate of the two types are shown in Figure 1. The distribution in the case of `Movie` is highly skewed, with the top predicate accounting for more than 30% of the statements. In the case of `Book` we see a more uniform spread of statements across the predicates.



(a) `Movie` (b) `Book`

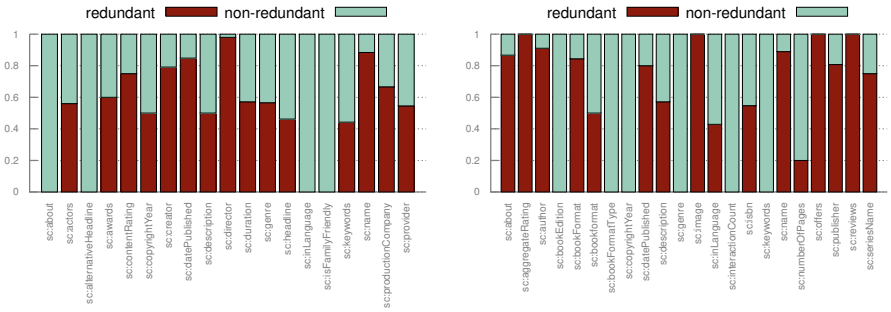
Fig. 1: Statement distribution across predicates for types `Movie` and `Book`.

In terms of *redundancy*, we notice that the amount of redundant information varies across predicates. Figure 2 shows the proportion of redundant information for the different predicates for the two entity types. It is worth noting that for *Object* predicates (e.g. `sc:actor`) we consider a statement to be duplicate if the corresponding `sc:label` are the same, whereas in the case of *Datatype* predicates we simply match the corresponding values.

[4] `https://lucene.apache.org`

We note that due to the inherently different nature of the types `Movie` and `Book`, redundant information is emphasized for different predicates and with a varying degree of redundancy. This is expected given that the *range* of these predicates for different types varies. For example, for type `Movie`, we note that predicate `sc:director` has a high degree of redundancy since the *sc:director* of an entity of type `Movie` is a very frequently provided i.e. important property. Hence it's occurring very often in the result set (and naturally is the same).

In terms of *non-redundancy*, we note that the highest degree of information comes from predicates whose range are *literals*. Specifically, `sc:headline` has the highest degree of non-redundancy with 53%, `sc:inLanguage` with 57% degree of non-redundancy, for type `Movie` and `Book`, respectively.



(a) `Movie` (b) `Book`

Fig. 2: Redundancy of information across predicates for types `Movie` and `Book`.

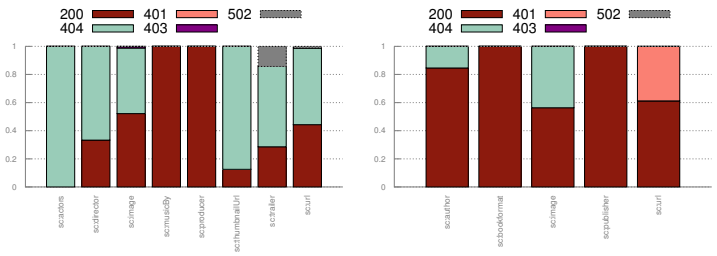
Information Inconsistency. Meusel and Paulheim [1] provide a preliminary analysis on some of the most common errors encountered in markup data. The errors range from syntactical to semantic errors, i.e. typos or misinterpretation and misuse of vocabulary terms. In this work, we follow their guidelines and measure such violations.

In terms of *semantic* errors, we focus on the violation of *Object* and *Datatype* properties. On average we find up to 3% and 4% semantic errors for `Movie` and `Book`, respectively. Other errors like *range* violations, we focus on few predicates which have specific requirements on the literal value, like `datePublished` and find 20% and 45% wrong values for `Movie` and `Book`, respectively.

Information Accessibility. An interesting aspect we note in Figure 1 is the high proportion of statements that contain links to external resources. Considering that such statements account for approximately 12% of markup data (on average for both types), we assess the accessibility of such links considering two aspects: (i) *HTTP response* from the URLs (i.e. HTTP Status Code=200 in case the URL is accessible), and (ii) the *content type* (e.g. *text/html* etc.)

In Figure 3 we show the availability of such URLs for both types. We note that there is a significant difference between the entity types, where for type `Movie` the ratio of unavailable links is higher in comparison to `Book`.

In terms of *content-type*, the majority of URLs is of content–type *text/html*. In the case of `Movie` we have 65% *text/html* and 35% *jpg/png*, similarly for `Book`, we have 70% *text/html* and 30% *jpg/png*. It is worth noting that even though there is a high amount of information available in the form of markup data, for tasks like entity retrieval or entity summarization, the available information presents a non-negligible set of resources which cannot be ignored.



(a) `Movie` (b) `Book`

Fig. 3: HTTP status codes for the different predicates for type `Movie` and `Book`.

3 Conclusion

In this work, we presented a preliminary analysis of some of the attributes of Web Markup data, and the challenges associated with them. While obtaining a ground truth for marked up facts is costly, we focused our investigation on a preliminary subset of the Web Data Commons.

When attemptng tasks like *entity retrieval*, which we envisage as one of the prerequisites for more advanced tasks such as knowledge base augmentation, markup data poses several key challenges. Our study underlines that there are inherently different characteristics of markup data when compared to traditional Linked Data and knowledge graphs, where characteristics even vary heavily across different entity types.

Some of our preliminary findings include (i)Statements about the same entities sourced from different PLDs, such as description, keywords, ratings etc., usually contains complementary information traditional Linked Data and knowledge bases (e.g. DBpedia); (ii)Predicates that encode *dates* usually contain a large amount of errors, and thus require plausability checks and further quality assurance metrics; (iii) When considering standard IR indexes, retrieval beyond top–50 does not provide any additionally relevant information in most cases for entity-centric queries. While these findings are based on a small subset of markup data, focused on two specific types, related case studies (see [5]or [4] have uncovered similar challenges in other domains.

References

1. R. Meusel and H. Paulheim. Heuristics for fixing common errors in deployed schema.org microdata. In *ESWC*, volume 9088, pages 152–168, 2015.
2. R. Meusel, P. Petrovski, and C. Bizer. The webdatacommons microdata, rdfa and microformat dataset series. In *The Semantic Web–ISWC 2014*, pages 277–292. Springer, 2014.
3. D. Ritze, O. Lehmberg, Y. Oulabi, and C. Bizer. Profiling the potential of web tables for augmenting cross-domain knowledge bases. In *WWW*. ACM, 2016.
4. P. Sahoo, U. Gadiraju, R. Yu, S. Saha, and S. Dietze. Analysing structured scholarly data embedded in web pages. Apr. 2016.
5. D. Taibi and S. Dietze. Towards embedded markup of learning resources on the web: An initial quantitative analysis of lrmi terms usage. In J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, and B. Y. Zhao, editors, *Companion Proceedings of the 25th International Conference on World Wide Web (WWW2016)*, pages 513–517. ACM, 2016.
6. R. Yu, U. Gadiraju, X. Zhu, B. Fetahu, and S. Dietze. Entity summarisation on structured web markup. In *ESWC: Satellite Events*, 2016.