Towards Entity Summarisation on Structured Web Markup

Ran Yu, Ujwal Gadiraju, Xiaofei Zhu, Besnik Fetahu and Stefan Dietze

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

Abstract. Embedded markup based on Microdata, RDFa, and Microformats have become prevalent on the Web and constitute an unprecedented source of data. However, RDF statements extracted from markup are fundamentally different to traditional RDF graphs: entity descriptions are flat, facts are highly redundant and granular, and co-references are very frequent yet explicit links are missing. Therefore, carrying out typical entity-centric tasks such as retrieval and summarisation cannot be tackled sufficiently with state of the art methods. We present an entity summarisation approach that overcomes such issues through a combination of entity retrieval and summarisation techniques geared towards the specific challenges associated with embedded markup. We perform a preliminary evaluation on a subset of the Web Data Commons dataset and show improvements over existing entity retrieval baselines. In addition, an investigation into the coverage and complementary of facts from the constructed entity summaries shows potential for aiding tasks such as knowledge base population.

Keywords: Entity Summarisation, Web Data Commons, Fact Selection

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 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[4]. Considering the upward trend of adoption - the proportion of pages containing markup increased from 5.76% to 30%

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

[2] http://www.w3.org/TR/microdata

[3] http://microformats.org

[4] http://www.webdatacommons.org

between 2010 and 2014 - and the still comparably limited nature of the investigated Web crawl, the scale of the data suggests potential for a range of tasks, such as entity retrieval, knowledge base population, or entity summarisation.

However, facts extracted from embedded markup have different characteristics when compared to traditional knowledge graphs and Linked Data. Co-references are very frequent (for instance, in the WDC2013 corpus, 18,000 entity descriptions of type `Product` are returned for query `'Iphone 6'`), but are not linked through explicit statements. In contrast to traditional strongly connected RDF graphs, RDF markup statements mostly consist of isolated nodes and small subgraphs. In addition, extracted RDF markup statements are highly redundant and often limited to a small set of predicates, such as `schema:name`. Moreover, data extracted from markup contains a wide variety of syntactical and semantic errors, ranging from typos to the frequent misuse of vocabulary terms.

These distinctive characteristics highlight the challenges when aiming to summarise entities sourced from embedded markup. Initial works such as the *Glimmer* search engine[5] have applied traditional *entity retrieval* techniques [1] to embedded markup (WDC corpus). However, given the large amount of flat and highly redundant entity descriptions, practical use of search results obtained in that way is limited. One major issue with such approaches are the unresolved entity co-references and entity fact redundancies. Therefore, applying *entity summarisation* techniques in order to obtain a homogeneous entity summary, assembled from all extracted facts seems the most promising approach in order to answer entity-centric queries. In this work we present an entity summarisation approach building on established entity retrieval approaches for obtaining a set of candidate entity descriptions and combined with algorithms for selecting and ranking distinct facts based on the clustering of candidate facts.

2 Approach

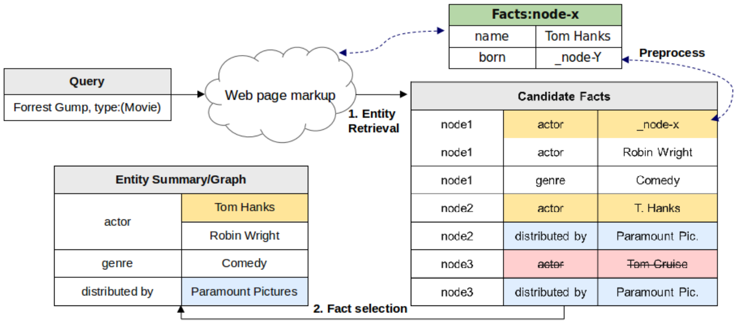
An entity summary consists of a set of facts, i.e. $\langle$ predicate, object $\rangle$ pairs. Given an entity-centric query, we generate the corresponding entity summary through a two-step approach: (i) entity retrieval, and (ii) fact selection. Figure 1 shows the summarisation process for query `Forrest Gump, type:(Movie)`.

Entity Retrieval. A prerequisite to construct entity summaries is the entity retrieval process. In our case, entities consist of a collection of facts $\langle$ predicate, object $\rangle$. For this purpose, we build a standard IR index, where instead of documents we consider entity descriptions (set of facts associated with an entity). Next, we retrieve entity descriptions based on the BM25 retrieval model.

A necessary step to further improve the entity retrieval step is to resolve the *object properties* in such entity descriptions, i.e., $\langle s_1$, `schema:actor`, $s_2\rangle$. We rely on a simple heuristic, where if such object values correspond to another entity in the WDC dataset, we replace it with it literal label `schema:name` in s_2.

Clustering-Based Fact Selection(CBFS). In this step, we select facts from the entity descriptions retrieved in the previous step. Our approach is restricted to facts associated with predicates from *schema.org*.

[5] http://glimmer.research.yahoo.com/



Facts:node-x	
name	Tom Hanks
born	_node-Y

Candidate Facts		
node1	actor	_node-x
node1	actor	Robin Wright
node1	genre	Comedy
node2	actor	T. Hanks
node2	distributed by	Paramount Pic.
node3	~~actor~~	~~Tom Cruise~~
node3	distributed by	Paramount Pic.

Entity Summary/Graph	
actor	Tom Hanks
	Robin Wright
genre	Comedy
distributed by	Paramount Pictures

Fig. 1: Overview of the entity summary pipeline.

Clustering. One major issue to address in the fact selection process, is canonicalizing different surface forms corresponding to a specific entity, e.g, `Tom Hanks` and `T. Hanks` are equivalent surface forms representing the same entity. To find duplicates and near duplicates, we first cluster the associated values at the predicate level into n clusters $(c_1, c_2, \cdots, c_n) \in C$. In this way, based on string similarity metrics we can canonicalize equivalent surface forms. For the clustering process we employ the X-means algorithm [3], which automatically determines the amount of clusters.

Fact Cardinality. Another challenge we address is the cardinality of predicates. Depending on the predicate, the number of correct statements varies. For example, `schema:actor` is associated with multiple values, whereas `schema:duration` normally has only one valid statement.

Fact Selection. Assuming that false entity facts have lower frequency, we eliminate irrelevant facts by choosing facts that are closer to the cluster's centroid and further meet the following criteria:

$$|c_j| > \beta \cdot \max(|c_k|), c_k \in C \tag{1}$$

where $|c_j|$ denotes the size of cluster c_j, and β is a parameter used to adjust the number of facts. In our experiments, β is empirically set at 0.5.

3 Experimental Setup and Evaluation

Dataset and Queries. We use a subset of the WDC 2014 dataset for experiments (entities of type `movie`), which contains 77 million RDF quads. We randomly select entities of type *movie* from Wikipedia as queries. Next, we keep only those queries that have at least 50 retrieved entities and have a high BM25 score (higher than 8). This is to ensure the correctness of entity summaries. Finally, we are left with 26 queries for use in our evaluation.

Performance. We consider BM25 as the baseline, and take the top 50 distinct facts from the retrieved entities, where the facts are ranked based on the associated rank to the original entity descriptions.

We use crowdsourcing to identify the precision of the retrieved facts for each each entity-centric query. Five different crowd workers provide us with binary labels for each fact (*correct*, *incorrect*) which are used as a ground truth. In the end, we get labels of 4901 distinct fact, which is also the candidate set of our CBFS approach.

Furthermore, we measure the diversity of the finally constructed entity summary, as the number of distinct *correct* facts for each predicate in our summary.

Table 1 shows the evaluation results. Our final approach, when compared to the standard IR approach, has a gain of 5.5% in terms of precision. Additionally, due to the clustering module in our approach, we have 12.5% more facts at the top 50 cutoff when compared to the baseline. This however, is intuitive given that the standard IR does not perform any de-duplification process.

Table 1: Performance of the proposed entity summarization approach.

Approach	Fact	Correct	*P*	Distinct	*Dist*%
CBFS	1075	895	**0.833**	876	**97.9**
BM25@50	1124	877	0.778	749	85.4

Coverage Gain. To evaluate the potential of our approach for aiding knowledge base augmentation tasks, we measure the coverage gain by comparing our results to DBpedia. We use *coverage gain (CG)*, i.e. the percentage of facts detected that are not available in DBpedia. We manually compare the fact selection results with corresponding DBpedia resources to determine the CG.

We found that 57% of the facts detected by our approach do not exist in DBpedia. Some of the facts correspond to new predicates. We also calculate the extra coverage considering only the predicates that exist in DBpedia, and the CG is 33.4%. This suggests that WDC dataset provides richer and diverse information in comparison to DBpedia and our approach as well as markup data in general are able to complement the knowledge available in DBpedia.

4 Conclusion and Future Work

We have introduced an entity summarisation approach. Our experimental results suggest potential for exploiting Web markup as a novel resource for tasks such as entity summarisation or knowledge base population. While our current approach is based on a simplistic pipeline involving a range of heuristics, as part of current and future work we are working on a fully automated approach for fact selection. In addition, more focused experiments aim at investigating the performance of our approach for specific knowledge base population tasks.

References

1. R. Blanco, P. Mika, and S. Vigna. Effective and efficient entity search in rdf data. In *The Semantic Web–ISWC 2011*, pages 83–97. Springer, 2011.
2. R. Meusel, P. Petrovski, and C. Bizer. The webdatacommons microdata, rdfa and microformat dataset series. In *The Semantic Web–ISWC 2014*. Springer, 2014.
3. D. Pelleg, A. W. Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, pages 727–734, 2000.