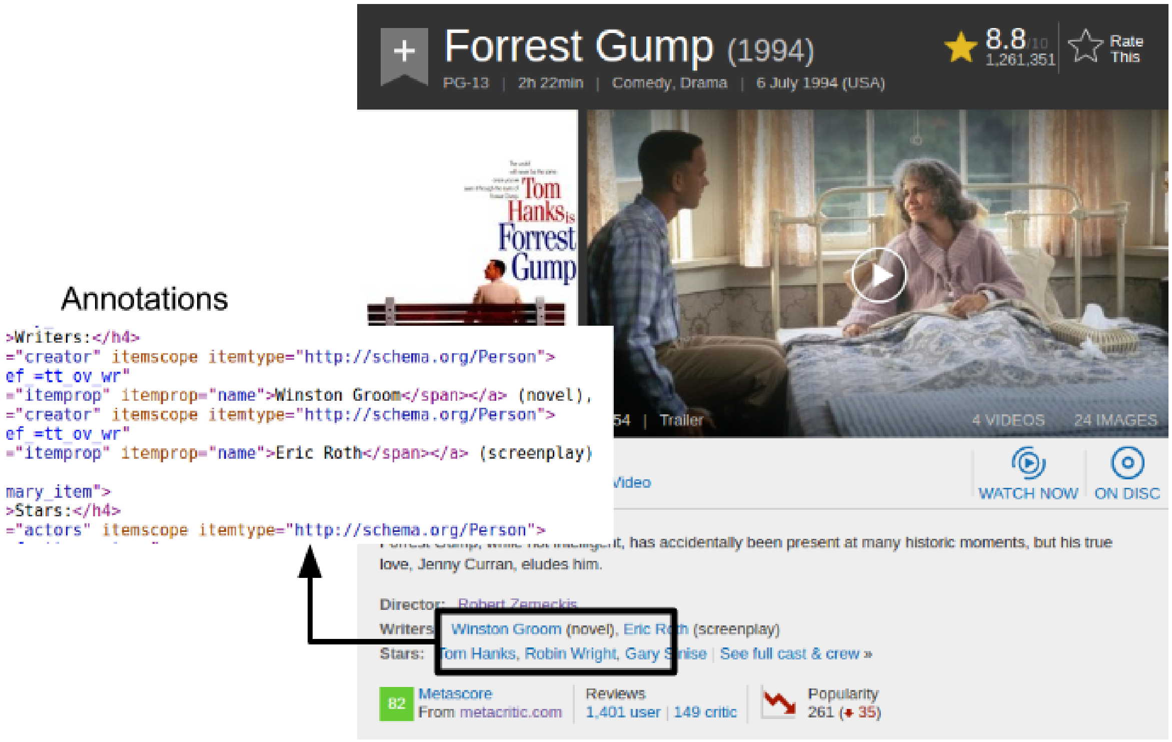
Clustering of Entities on the Web

Entity Resolution & Clustering with Weighted Structural Similarity Metric and Pruning

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



Annotations

```
>Writers:</h4>
="creator" itemscope itemtype="http://schema.org/Person">
ef_=tt_ov_wr"
="itemprop" itemprop="name">Winston Groom</span></a> (novel),
="creator" itemscope itemtype="http://schema.org/Person">
ef_=tt_ov_wr"
="itemprop" itemprop="name">Eric Roth</span></a> (screenplay)

mary_item">
>Stars:</h4>
="actors" itemscope itemtype="http://schema.org/Person">
```

Forrest Gump (1994)

PG-13 | 2h 22min | Comedy, Drama | 6 July 1994 (USA)

8.8/10 1,261,351

Forrest Gump, while not intelligent, has accidentally been present at many historic moments, but his true love, Jenny Curran, eludes him.

Director: Robert Zemeckis
Writers: Winston Groom (novel), Eric Roth (screenplay)
Stars: Tom Hanks, Robin Wright, Gary Sinise | See full cast & crew »

The Web Data Commons (WDC), 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 from the Common Crawl.

Considering the upward trend of adoption - the proportion of pages containing markup increased from 5.76% to 30% 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, entity summarization* and *knowledge base augmentation*.

Due to the nature of WDC, co-references are very frequent (for instance, in the WDC 2013 corpus, 18,000 entity descriptions of type *Product* are returned for query '*Iphone 6*', but are not linked through explicit statements. Finding the entity descriptions correspond to the same real-world entity is a prerequisite for the applications mentioned above

Although different similarity measures are adopted in previous entity resolution works [1], the role of semantics is still ignored to a large extent. That is, different properties should have varied contribution when computing the similarity. A movie has several properties, eg "name" and "actor", where "name" is a stronger indicator of similarity. That is, two entities of type movie having the same or similar names are highly similar, while the same actor does not necessarily indicate strong similarity. Thus we aim at developing an entity similarity measurement approach that can automatically distribute weight across properties and choose different similarity metrics with regards to the datatype/range of the property. Furthermore, consider the scale of data to process,e.g.there are 168,233,105 entities of schema.org type *Person* alone are contained in the WDC 2014 dataset, pairwise comparison on data of such scale would not be a sufficient approach, so pruning(or so-called blocking for entity-coreference resolution) techniques are also required.

References

[1] Chapter 2: Christophides, Vassilis, Vasilis Efthymiou, and Kostas Stefanidis. "Entity Resolution in the Web of Data." Synthesis Lectures on the Semantic Web 5.3 (2015): 1-122.
[2] Ristoski, Petar, and Peter Mika. "Enriching Product Ads with Metadata from HTML Annotations." International Semantic Web Conference. Springer International Publishing, 2016.
[3] Volz, Julius, et al. "Silk-A Link Discovery Framework for the Web of Data."LDOW 538 (2009).

M.Sc./Diploma thesis project

L3S Research Center offers a M.Sc. project. Taking into account the research context described above, the aim of the thesis is to develop an approach to cluster the entity descriptions in a large scale structured data dataset extracted from the web markup.

Are you interested in working with data mining, semantic web and search technologies? Want to be part of an international research team working with a new exciting research topic? The research tasks would entail the following:

- Research state of the art in entity coreference resolution
- Blocking / pruning technique to improve the computing efficiency
- Measuring entity similarity, which includes weighting the different property, and choosing different metric for different datatype
- Develop scalable methods to cluster the web markup dataset

You should be:

- An interested and motivated worker with a keen will to learn
- Familiar with information retrieval system
- Familiar with basic data mining algorithms
- Familiar with Linked Data
- Familiar with programming languages (ideally Python, Java)

Are you interested or have questions?

Contact us:
Dr. Stefan Dietze, Email: dietze@l3s.de,
Ran Yu, Email: yu@l3s.de,

Forschungszentrum L3S, Appelstr. 4, 30167 Hannover