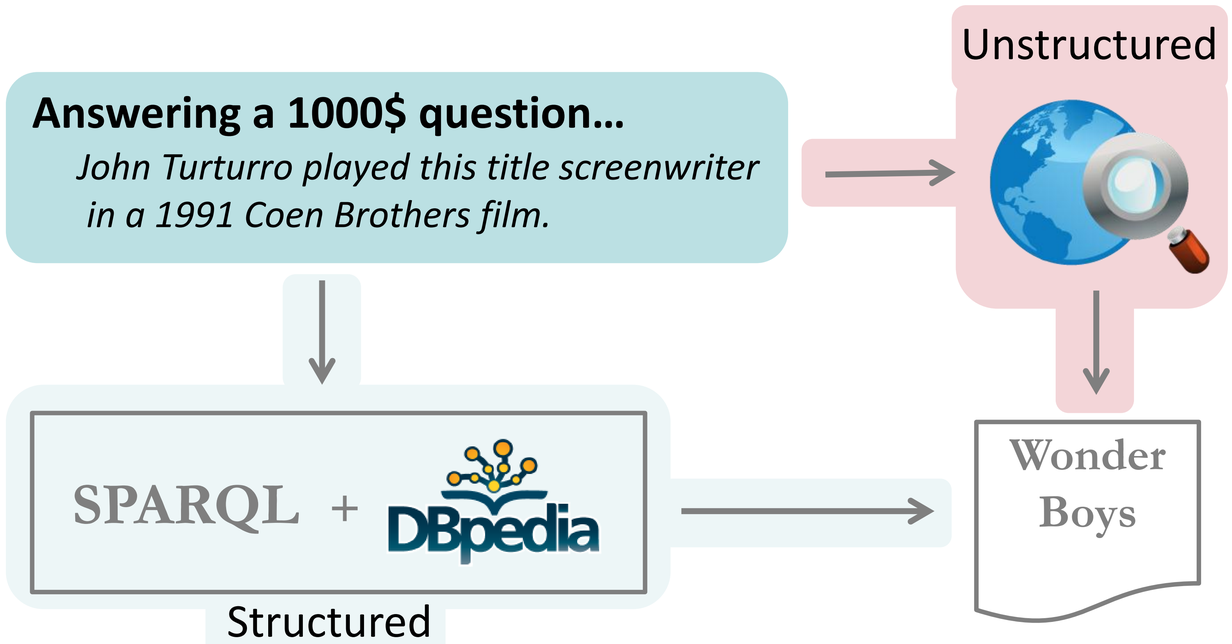
Design and Evaluation of an IR-Benchmark for SPARQL Queries with Fulltext Conditions

Arunav Mishra, Sairam Gurajada, and Martin Theobald

5th International Workshop on Exploiting Semantic Annotations in Information Retrieval, CIKM 2012



Two worlds of querying:

Unstructured

+ Easy to formulate (no joins)
- Misses user intention

Structured

+ Better captures user intention and unambiguous
- Need expertise on language and schema

Unify(DBpedia+YAGO, WIKIPEDIA) ⇒ WIKI-LOD Benchmark

FTContains operator for SPARQL

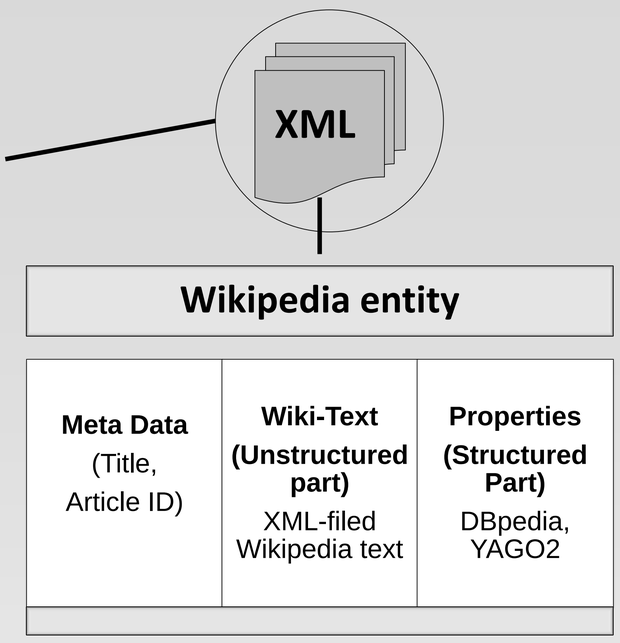
- to add keyword constraints to structured query

INEX Initiative for the Evaluation of XML retrieval

2012
Linked Data Track

Wiki-LOD Benchmark collection[1]

Property	Count
XML Documents	3,164,042
Wikipedia Category Articles	266,134
Wikipedia Entity Articles	2,053,050
Wikipedia Entity Articles with Infoboxes	907,304
Other Wikipedia Articles	844,857
Resolved DBpedia Links	36,941,795
Resolved YAGO2 Links	32,941,667
Intra-Wiki Links	22,235,754
External Web links	7,214,827
Imported DBpedia Properties	168,374,863
Imported YAGO2 Properties	23,634,511




```
<lodxml xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xml:lang='en'
xmlns:xhtml='http://www.w3.org/1999/xhtml' encoding='UTF-8'>
   <article title='Albert_Einstein'>
      <wikipedia>
         <template type='Metadata'>                          Meta Data
            <tag name='id'>736</tag>
            <tag name='title'>Albert_Einstein</tag>
         </template>……
         <infobox type='person'>                          Infobox Property
            <tag name='name'>Albert Einstein</tag>
            <tag name='image'>Einstein 1921 portrait2.jpg</tag>
         </infobox>……..
          <paragraph>Albert Einstein was born in<space/>
           <link>                                                  LOD Links
            <wikilink href='http://en.wikipedia.org/wiki/Ulm'>Ulm</wikilink>
            <dbpedia href='http://dbpedia.org/resource/Ulm'/>
            <yago ref='Ulm'/>
            </link>……..
      </wikipedia>
     <dbpediaproperties>                              DBpedia and Yago Properties
         <property name='http://dbpedia.org/property/birthPlace'>
            <object name='http://dbpedia.org/resource/German_Empire'/>
         </property>……..
     </dbpediaproperties>
     <yagoproperties>
         <property name='isKnownFor'>
            <object name='Einstein_field_equations'/></property>
     </yagoproperties>
   </article>
</lodxml>
```

A Sample "Albert Einstein" Wiki-LOD XML Document

Query Benchmark[1]

90 handcrafted SPARQL queries with fulltext conditions:

- 50 jeopardy queries having one target entity as answer
- 40 natural language queries returning ranked list of one or more entities

Example query*: John Turturro played this title screenwriter in a 1991 Coen Brothers film*

```
select ?s where {
?s <http://dbpedia.org/ontology/starring> ?x.
?s <http://dbpedia.org/ontology/director> ?y.
FILTER FTContains (?x,"John Turturro").
FILTER FTContains(?y,"Coen brothers").
FILTER FTContains(?s,"John Turturro played 1991 Coen Brothers film"}
```

Setup and Evaluation

- Manually translated 90 queries into SPARQL with fulltext conditions.
- Built a fulltext index (using BM25 ranking model) over the textual components of Wiki-LOD documents.
- Assessed 50 of the Jeopardy queries using Amazon Mechanical Turk

Novelties & Goals

- Benchmark collection and queries help to bring together two worlds (structured & unstructured) retrieval models and challenge new types of indexing and querying
- Linked Data pointers (Yago and DBpedia) in Wiki-LOD allow people to extend into almost arbitrary RDF (LOD) collections
- The new collection may be of high interest for other IR applications, like the clustering, classification of entities, summarization, taxonomy building, and many more

References

[1] INEX 2012 Linked Data Track.https://inex.mmci.uni-saarland.de/.
[2] Media Wiki. http://en.wikipedia.org/wiki/MediaWiki/.
[3] C. Bizer et al. DBpedia - A crystallization point for the Web of Data. J. Web em., 7(3), 2009.
[4] J. Hoffart et al. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. AI '12

MAX-PLANCK-GESELLSCHAFT
max planck institut informatik
{amishra, gurajada, mtb} @ mpi-inf.mpg.de
Max-Planck Institute for Informatics, Saarbrücken, Germany