Extracting Time Information from YAGO

Presented to:

Professor Abdullah Tansel

By:

Hussein Ghaly

Abstract

YAGO is a Knowledge Base that includes millions of structured facts about people and things. These facts are temporally and spatially enhanced, such that each fact has an associated (meta) fact indicating when/where this fact was valid. The main objective of the current project was to do reification of these facts, so that for each fact that has some meta facts corresponding to it, we combine all these associated meta facts and sort them. This was achieved using Python database object “shelve”, with the keys being the fact ID for easy retrieval and grouping.

Introduction

Semantic data is very useful in allowing computers and devices to process real-world facts for a variety of applications, and has a potential to improve the search output, and currently we see Google is starting to include some semantic elements within its search results. There are several semantic data sources, which rely mainly on RDF triplets, that express any fact by three things (Subject, Predicate, and Object). Examples of such sources are DBPedia, Freebase and YAGO. YAGO is built automatically from Wikipedia, GeoNames, and WordNet, with a total number of facts: 158,948,016. In this project, YAGO was chosen to investigate the how facts are linked together in a way that one fact can be a subject of another fact, in order to be able to indicate more information about the fact.

More about YAGO can be found in the following paper:
http://people.mpi-inf.mpg.de/~kberberi/publications/2010-mpii-tra.pdf

Data

Simple Facts:

They are in the form:

<entity1_id> <relationship> <entity2_id>

So it is basically a list of facts that describe the relationship between a pair of entities.

Literal Facts:

They are in the form:

<entity1_id> <relationship> <string>

The string can be a date or a number or anything that does not necessarily represent an entity, but rather an attribute of entity1.

Meta Facts:

<fact_id> <relationship> <string>

This is probably the major contribution of YAGO, creating these meta facts which are basically facts about facts, and they can be very useful to describe various facts that are linked together, such as a certain person holding a political position, so we would want to know from what time till what time. It essentially adds the special and temporal dimension to facts.

Schema:

It indicates the different relationships and how they are organized.

Statistics:

It provides a count for each type of relationships.

Method

There are three main steps in this process:

1. create_facts_shelve.py: This is the first step of the process, we create a shelve object for all the facts from both yago facts (yagoFacts.tsv) and yago literal facts (yagoLiteralFacts.tsv). The output is the shelve file (updated_facts.shelve) which has its keys the fact IDs.
2. yago_group_reified.py: This is the second step of the process. This code uses the facts that are in the shelve file (updated_facts.shelve) and the yago meta facts (yagoMetaFacts.tsv) to group all the meta facts related to a certain fact together. Its output is the file (updated_reified.shelve), which is a shelve file, its keys are the simple fact IDs, and its values are the meta facts associated with these simple facts
3. yago_reified_out.py: This is the third step of the process. It uses the facts shelve (updated_facts.shelve) and the reified shelve (updated_reified.shelve) to create a text file that combines each fact with its reified meta facts.

Statistics

The following is a break down for which time relationships occurred most frequently. The most frequent was <wasBornOnDate>.

<wasBornOnDate>	0.00506828597345
<endedOnDate>	6.29136509637e-09
<diedOnDate>	0.00227705264343
<startedOnDate>	5.66222858673e-08
<wasCreatedOnDate>	0.00454649273508
<occursSince>	0.00347991131893
<occursUntil>	0.00212096387538
<happenedOnDate>	0.00131509033746

<wasDestroyedOnDate>	0.000276662780113

Ratio of time facts to all facts: 0.0190845225775

Conclusion

The main goal of this project was to better understand time data stored in YAGO by investigating the reification of simple and literal facts. The main contribution added was the ability to create a simple database object (python shelve) that allows accessing and combining the facts easily according to their ID, this can be extended through some web service to use this database object to answer sophisticated queries. Possible extensions of this project can include processing other structured data sources (freebase, DBPedia… etc) in order to combine their facts with the ones from YAGO, thus extending the knowledge base and add more credibility to the facts included.