WikipediaBase Architecture

Chris Perivolaropoulos

Sunday 21 February 2016

Contents

1 Infobox

Infoboxes are tables that are commonly used in wikipedia to provide an overview of the information in an article in a semi structured way. Infoboxes are the main source of information for WikipediaBase.

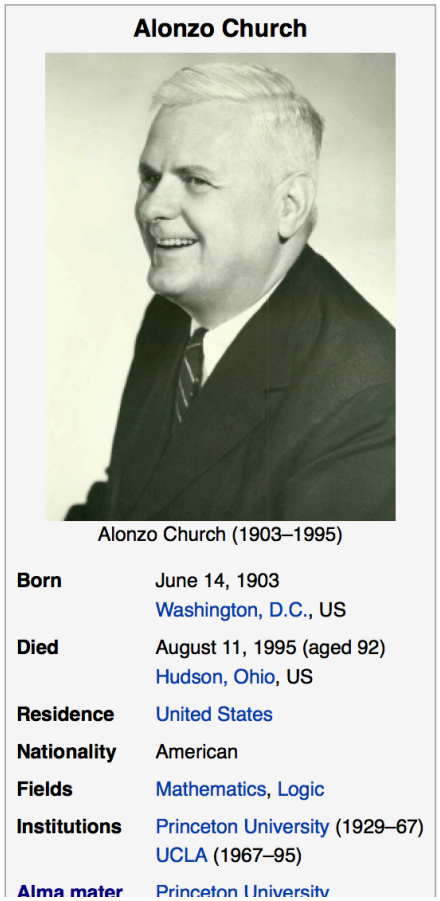

Figure 1: An example of an infobox

In mediawiki markup terms an infobox is a markup template with a type that gets rendered into html so that the provided information makes sense in the context that it is provided. For example:

```
{{Infobox scientist
| name              = Gerhard Gentzen
```

```
| image              = Gerhard Gentzen.jpg
| image_size         =
| alt                =
| caption            = Gerhard Gentzen in Prague,
   1945.
| birth_date         = {{Birth date|1909|11|24}}
| birth_place        = [[Greifswald]], [[Germany]]
| death_date         = {{Death date and age
   |1945|8|4|1909|11|24}}
| death_place        = [[Prague]], [[Czechoslovakia]]
| nationality        = [[Germany|German]]
| fields             = [[Mathematics]]
| workplaces         =
| alma_mater         = [[University of Gottingen]]
| doctoral_advisor   = [[Paul Bernays]]
| doctoral_students  =
| known_for          =
| awards             =
}}
```

will yield:

Figure 2: An example of an infobox

Infobox types are organized into a fairly wide hierarchy. For example Template:Infobox Austrian district is a special case of a Template:Infobox settlement and each is rendered differently. For our purposes, and to mirror the markup definition of infoboxes, an infobox I with attributes a_i and values v_i is a set of pairs (a_i, v_i) together with a infobox type t. Each attribute a_i and value v_i have two forms:

- a rendered form, a_i^r and v_i^r respectively, which is the rendered HTML representation and
- a markup form, a_i^m and v_i^m which is the mediawiki markup code that corresponds to them.

An article may have more than one infoboxes, for example Bill Clinton article has both Infobox Officeholder and Infobox President infoboxes.

The `Infobox` class is the basic data type for accessing information from the infobox of an article. `Infobox`, as well as `Article`, are what one would use were they to use wikipediabase as a python library. The methods provided by an infobox are:

types Because we retrieve an infobox based on a symbol name (ie page name), a single `Infobox` may actually be an interface for multiple infoboxes. There is a separate method, based on this one, for getting types in a format suitable for START.

Value access is possible provided either a_i^r or a_i^m.

Rendered keys are provided using the `MetaInfobox` (see below).

Infobox export to python types, namely:

- dict for $a_i^r \rightarrow v_i^r$ or $a_i^m \rightarrow v_i^m$
- the entire infobox rendered, or in markup form.

2 Infobox tree

3 MetaInfobox

The `MetaInfobox` is a subclass of the `Infobox` that provodes information about the infobox, most importantly a map between markup attributes. Say we have an infobox of type I which has attributes $a_1, ..., a_n$. Each instance of that infobox I defines

It is an infobox with all the valid attributes and each value is all the names of all attributes that are equivalent to them. Eg An infobox of type Foo that has valid attributes A, B, C and D and A, B and C are equivalent has a meta infobox that looks something like:

Attribute	Value
A	!!!A!!! !!!B!!! !!!C!!!
B	!!!A!!! !!!B!!! !!!C!!!
C	!!!A!!! !!!B!!! !!!C!!!
D	!!!D!!!

4 Article

The `Article` data structure is responsible for accessing any resource relevant to the article at large.

5 Fetcher

The fetcher is an abstraction over the communicatioln of WikipediaBase with the outside world. It is a singleton object that implements a specific interface.

Fetchers are organized in an inheriting hierarchiy

BaseFetcher The baseclass for fetchers, it will return the symbol instead of trying to resolve it in any way

Fetcher contains the core functionality of a a fetcher. It will fetch articles from *wikipedia.org*. It is possible to direct it to a mirror but wikipedia-mirror's runtime performance turned out to be prohibitive.

CachingFetcher inherits fetcher and retains it's functionality, only it uses Redis to cache the fetched symbols. It is the default fetcher for wikipediabase.

StaticFetcher is a class that implements the `BaseFetcher` interface but instead of reaching out to some data source for the data the return values are statically defined. It is used most notably by `MetaInfobox` to use the `Infobox` functionality to convey arbitrary information.

By default, markup is fetched from the backend. If $\text{force}_{\text{live}}$ is set to True, the markup will be fetched from live wikipedia.org

When tests are ran on TravisCI, we always want to use live data. We check if Travis is running tests by looking at the $\text{WIKIPEDIABASE}_{\text{FORCELIVE}}$ env variable.

6 Renderer

Renderers are singleton classes that are useful for rendering mediawiki markup into HTML. Originally the wikiepedia sandbox was used by wikipediabase for rendering pages because it is slightly faster than the API, but the wikipedia-mirror was really slow at this and wikipedia.org would consider it an abuse of the service and block our IP. For that reason we eventually switched to

the API with Redis caching, which works out pretty well because `Renderer` objects end up being used only my `MetaInfobox` which has quite a limited scope, making thus cache misses rarely.

7 Caching

TODO, check what alvaro did with this

8 Logging

9 Utilities

9.1 General

9.2 Database utilities

10 Pipeline

When resolving a query WikipediaBase employs a pipeline of modules to figure out what the best way to respond would be.

10.1 Frontend

WikipediaBase can be used as a library but it's primary function is as a backend to START. The communication between START and WikipediaBase is carried out over a plaintext telnet connection on port {port} using EDN-like sexpressions. The frontend handles the network connection with START, translates the received queries into calls to knowledgebase and then translate the knowledgebase response into properly formulated sexpressions that it sends back over the telnet connection.

10.2 Knowledgebase

The knowledgebase is the entry point to the rest of wikipediabase. It uses the Provider/Acquirer pattern to transaprently provide the frontend with arbitrary methods. Those methods are responsible for chosing whether we are to resort to classifiers or resolvers (or any other mechanism) for answering the query. Available classifiers and resolvers become accessible to the knowledgebase automatically using their base class.

10.3 Classifiers

Each classifier is a singleton that implements a heuristic for deducing a set of classes of an object. An object may inhibit zero or more classes. There are a couple classifiers available at the moment. Typically a classifier will only deduce whether an object actually inhibits a specific class or not but that is not necessary.

1. Term

   The `TermClassifier` simply assigns the `wikipedia-term` class. Wikipediabase only deals with wikipedia related information.

2. Infobox

   The `InfoboxClassifier` assigns to a term the classes of the infobox. For example Bill Clinton's page contains the infobox:

```
    {{Infobox president
|name          = Bill Clinton
|image         = 44 Bill Clinton 3x4.jpg{{!}}border
|office        = [[List of Presidents of the United States|42nd]] [[President of t
|vicepresident = [[Al Gore]]
|term_start    = January 20, 1993
|term_end      = January 20, 2001
|predecessor   = [[George H. W. Bush]]
|successor     = [[George W. Bush]]
|order1        = 40th and 42nd [[List of Governors of Arkansas|Governor of Arkansa
|lieutenant1   = [[Winston Bryant]]<br>[[Jim Guy Tucker]]
|term_start1   = January 11, 1983
|term_end1     = December 12, 1992
|predecessor1  = [[Frank D. White]]
|successor1    = [[Jim Guy Tucker]]
|lieutenant2   = [[Joe Purcell]]
|term_start2   = January 9, 1979
|term_end2     = January 19, 1981
|predecessor2  = [[Joe Purcell]] {{small|(Acting)}}
|successor2    = [[Frank D. White]]
|office3       = 50th [[Arkansas Attorney General|Attorney General of Arkansas]]
|governor3     = [[David Pryor]]<br>[[Joe Purcell]] {{small|(Acting)}}
|term_start3   = January 3, 1977
|term_end3     = January 9, 1979
```

```
|predecessor3  = [[Jim Guy Tucker]]
|successor3    = Steve Clark
|birth_name    = William Jefferson Blythe III
|birth_date    = {{birth date and age |1946|8|19}}
|birth_place   = [[Hope, Arkansas|Hope]], [[Arkansas]], [[United States|U.S.]]
|death_date    =
|death_place   =
|party         = [[Democratic Party (United States)|Democratic]]
|spouse        = {{marriage|[[Hillary Clinton|Hillary Rodham]]|October 11, 1975}}
|relations     = ''See [[Clinton family]]''
|children      = [[Chelsea Clinton|Chelsea]]
|parents       = [[William Jefferson Blythe, Jr.]]<br>[[Virginia Clinton Kelley]]
|alma_mater    = [[Edmund A. Walsh School of Foreign Service|Georgetown University
|religion      = [[Baptists|Baptist]] {{small|(formerly [[Southern Baptist Convent
|signature     = Signature of Bill Clinton.svg
|signature_alt = Cursive signature of Bill Clinton in ink
|website       = {{url|clintonlibrary.gov|Library website}}
}}
```

And therefore gets the class `wikipedia-president`.

3. Person

   `PersonClassifier` assigns the class `wikibase-person` using a few heretics in the order they are described:

   (a) Category regexes

   Use the following regular expressions to match categories of an article.

   - `.* person`
   - `^\d+ deaths.*`
   - `^\d+ births.*`
   - `.* actors`
   - `.* deities`
   - `.* gods`
   - `.* goddesses`
   - `.* musicians`
   - `.* players`
   - `.* singers`

(b) Category regex excludes

Exclude the following regexes.

- `\sbased on\s`
- `\sabout\s`
- `lists of\s`
- `animal\`

(c) Category matches

We know an article refers to a person if the page is in one or more of the following mediawiki categories:

- `american actors`
- `american television actor stubs`
- `american television actors`
- `architects`
- `british mps`
- `character actors`
- `computer scientist`
- `dead people rumoured to be living`
- `deities`
- `disappeared people`
- `fictional characters`
- `film actors`
- `living people`
- `musician stubs`
- `singer stubs`
- `star stubs`
- `united kingdom writer stubs`
- `united states singer stubs`
- `writer stubs`
- `year of birth missing`
- `year of death missing`

For example Leonardo DiCaprio's page has the following categories:

- Leonardo DiCaprio
- 1974 births

- **Living people**
- 20th-century American male actors
- 21st-century American male actors
- American environmentalists
- American film producers
- American male child actors
- American male film actors
- American male soap opera actors
- American male television actors
- American people of German descent
- American people of Italian descent
- American people of Russian descent
- American philanthropists
- Best Actor AACTA Award winners
- Best Actor Academy Award winners
- Best Drama Actor Golden Globe (film) winners
- Best Musical or Comedy Actor Golden Globe (film) winners
- California Democrats
- Film producers from California
- Formula E team owners
- Male actors from Hollywood, California
- Male actors from Palm Springs, California
- Male actors of Italian descent
- People from Echo Park, Los Angeles
- Silver Bear for Best Actor winners

As it is obvious the list of categories is arbitrary and very far from complete. Multiple methods have been considered for fixing this. Some of them are:

- Supervised machine learning methods like SVM using other methods of determining person-ness to create training sets.
- Hand-pick common categories for person articles determined again with the other criteria

10.4 Resolvers

Resolvers are also singletons but their purpose is to find the value of the requested property.

1. Error
2. Infobox
3. Person
4. Section
5. Term

   If the article matches one or more of the following categories:

10.5 Lisp types

Lisp type instances are wrappers for python objects or values that are presentable in s-expression form that START can understand. They are created either from the raw received query and unwrapped to be useful to the pipeline, or by the answer WikipediaBase comes up with and then encoded into a string sent over telnet to START.