Category/Topic Age via Wikipedia

Sage Hahn
University of Vermont
Complex System Center
sahahn@uvm.edu

Thayer Alshaabi
University of Vermont
Complex System Center
talshaab@uvm.edu

1. Introduction

Projects, as they tend to be, can be said to be concerned with the exploration of previously ineffable notions - ours is no exception. Of particular interest are those prepended with the word "final", of which again ours, implicitly albeit, is no exception. Obfuscation aside, we sought to undertake the previously nebulous (now arguably at least 50% tangible) task of assigning a Wikipedia (Topic $\vee$ Category) an age [2] . Now we believe is as good a time as any to mention that for the citizens of Wikipedia (of which (Topic $\vee$ Category)'s are just one among many, *e.g*. dead presidents, notable hotels, meta lists) the concept of age functions differently then the simple integer representation we are used to. More different still then its alternate definitions (the dreaded verb and/or reference to a notable period of time). Age is instead vastly more complex, an entity in itself. Thus, as with attempting to explain any complex system, we must resort to descriptors of its behavior, analysis of its components, and of course wild guesses surrounding any underlying mechanisms which may or may not describe the observed behavior. In other words, we define a (Topic $\vee$ Category)'s age to represent the distribution of years associated with a page, and branch pages.

2. Methods

Within this section we present details surrounding the scraping tool developed for this project. Wikipedia, while relatively uniform in comparison to the rest of the wild web, presents an interesting 'target' for web scraping. In particular, a number of design decisions should be taken into account during interpretation of later sections, as they frame any semantic meaning introduced in our definitions of Age, Connections, Categories, *etc*.

We define at the highest level of interest, and most loosely, "Categories" as representative, typically, of academic disciplines. That being said, certain Categories lend themselves more readily towards our proposed analysis such as Mathematics, Chemistry and Physics in contrast towards more nebulous or highly specific topics such as History or Geodesy. Given a Category, we next define a 'Branch' as any sub topic within that larger category, say X, such that it would appear reasonably within a list titled "Branches of X" or "Fields of study of X", and importantly that it have an associated Wikipedia page.

For each Category considered we sought to collect the following information from both its associated Wikipedia page and those of its Branches.

1. The years associated with each item present within the References, Notes, Sources, Further Reading and all other similar lists present on a given page.

2. Links to people found within a page, where for each person found the following items are collected.

   - The year they were born and, if relevant, died.
   - All years mentioned across their whole page between when they were born and either when they died or 2019 (the present).

3. All internal Wikipedia links found within the body of the page, *i.e*. typically above the "See also" section and other lists of related links.

Towards this goal, we made significant efforts towards creating a comprehensive and robust tool, though due to the underlying and often idiosyncratic nature of text processing, edge cases understandably remain. While the full details surrounding our scraping procedure tend to be overly complicated, it bears mentioning some of our specific design decisions (*i.e*. we filtered out non-year numbers vs. here are 14 reg-ex patterns explained in excruciating detail, that we used to deal with bizarre edge cases).

Finding Branches for a given Category: Given a Category, potential Branches were found primarily through scraping the "Outline of CategoryName" page, which tends to contain lists of sub-fields. Notably, we did not make any effort to explicitly extract a hierarchical structure of branches here, instead treating each entry as equal (the idea

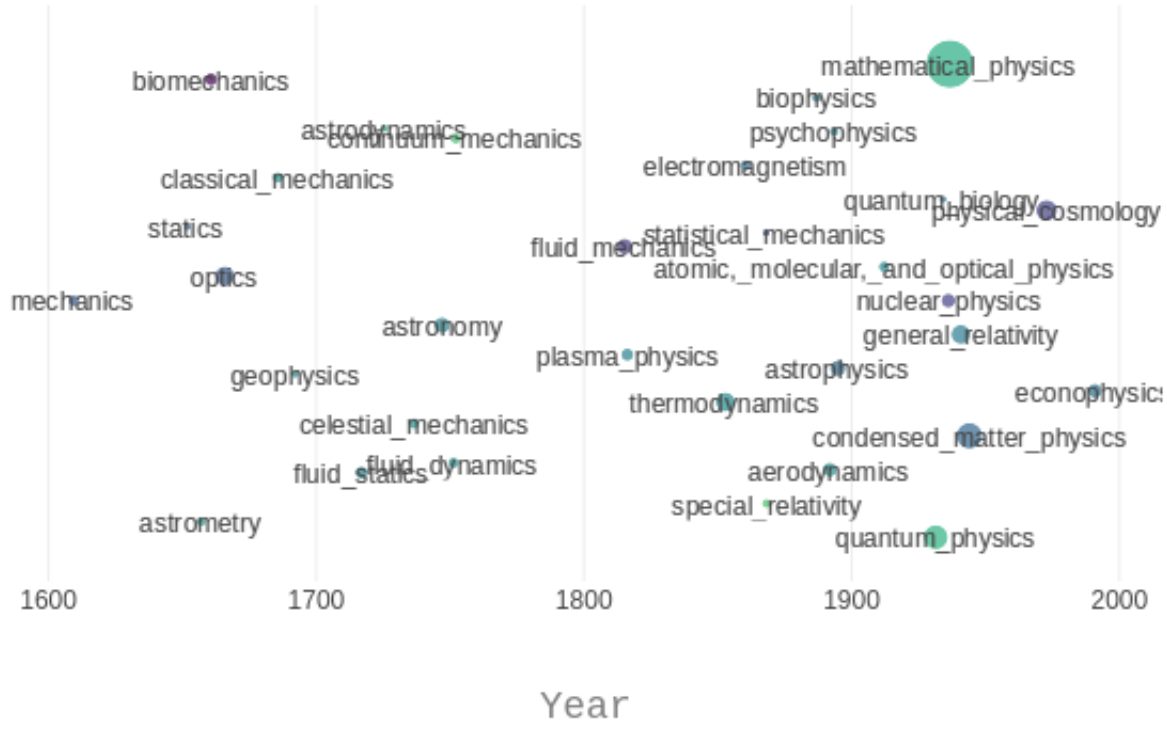

Figure 1: Branches of physics over time, where a branch is assigned a year based on the median of each person found within that branches mean year.

being, any structure could be discovered organically afterwards).

Determining if a page is about a person: Given a list of internal Wikipedia links, in order to distinguish if a given link represents a person we initially filter out non person links based on just name. Next, remaining links are visited and the corresponding html searched for revealing keywords. Specifically, we generated a list of keys found only on pages containing a person (*e.g.* id=Biography) and likewise a list of keys found only on pages not about a person (*e.g.* id=Definitions) in order to make a final decision.

Extracting born/died years from a persons page: We employed two separate methods in order accurately extract the year a person was born and if they died. Notably, this information if present within a persons page occurs either in the first sentence and/or within a dedicated 'vbox' on the side of the article. Both of these locations are scraped (with a number of inane corresponding rules, *e.g.* priority given to text within parentheses, higher digit numbers, checking for nearby B.C., other sanity checks *etc.*) and then combined with priority given to more explicit vbox years.

Extracting a year from a reference: In order to extract a year from a single note or reference item we preform arguably our most convoluted scheme, as there exist a huge number of different possible formats and edge cases. We begin by removing a number of known non-year numbers, either by locating indicators such as 'date retrieved' or through a number of regular expressions designed to filter out numbers present within a title or related field, *e.g.* "Chem491". Next, years are searched according to various priorities, for example first checking within brackets and then parentheses, along with considerations around length of digit found and if the reference contains a link to an outside url. Additionally, if an ISBN number is found, we attempt a look-up on that ISBN and if information is available extract year from there. If no year is found, the whole search procedure is repeated, but without removing the title field therefore making it a valid spot to search for year. Of additional consideration is the decision made that fields like 'date accessed' or 'archived date' would not be considered valid years, in this sense we are removing most static websites and vague references.

In this sense, we have built a (somewhat) robust tool capable of scraping a large variety of information from Wikipedia given a topic. All project code, as described above, can be found at

https://gitlab.com/thayerAlshaabi/wikiagenet.

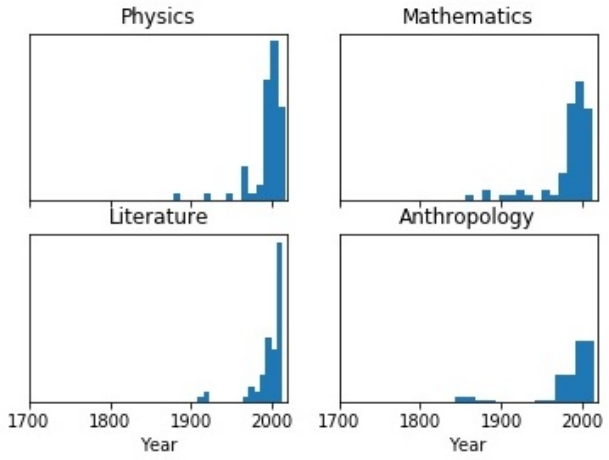

Figure 2: Normalized histogram of years found from references on a given categories Wikipedia page.

Topic	Age/Date	Topic	Age/Date
Comm.	01/28/1963	Literature	06/03/1834
Economics	04/05/1921	Pol. Sci.	09/12/1825
Psychology	04/02/1912	Mathematics	04/23/1817
Anthropology	03/30/1902	Neuroscience	06/03/1815
Comp. Sci.	08/14/1890	Linguistics	03/25/1804
Sociology	07/08/1888	Physics	11/02/1796
Chemistry	01/18/1877	Health Sci.	10/23/1760
Education	01/18/1870	Agriculture	01/03/1780
Engineering	10/15/1843	Library Sci.	06/24/1701
Biology	06/13/1839	Geology	06/09/1630

Table 1: A Topic's age/date represents the mean of all of the people's means, across all of the people found on all branches.

3. Results

3.1. References

Our first attempt to assign age relates directly the references found on a given topic. We initially scraped a number of example topics using the methodology described in Section 2. We collected the references and notes for each given topic. Years were then extracted from the raw reference element in the HTML body. Unfortunately, most recorded years (in the references section) do not date back far enough in time. Consequently, we ended up with boring plots where most citations for a given topic are clustered within the last century. Figure 2 shows a normalized histogram of references years found across a few example topics.

More importantly, it is almost impossible to find all relevant citations and references for a given topic within one Wiki page. In order to gather a more diverse collection of references, and as relevant references tend to be scattered across multiple pages, we scraped all years associated with all branches of any given topic. We then analyzed the new set of years collectively across each considered academic discipline hoping that a new paradigm would emerge. However, reference year results were very similar, suffering from the same problems as the former approach. In fact, the new set of reference year histograms were even more heavily skewed towards recent years (Why would two grad students ever have hopes and dreams?) (God save the queen!).

3.2. People

On the bright side, we managed to move on from the references catastrophe by introducing a new more meaningful approach. This time, we additionally analyzed the pages of all people mentioned within any given topic and/or branch. As discussed in section 2, while some sketchy methods were applied, our method yielded considerable accuracy in distinguishing person from non-person pages.

Ultimately, we extracted out all years mentioned within each person page (nuggets of year information). We then averaged out those years and assigned a year for each person to represent an 'active' year for that individual in our dataset. Going back to our original idea of assigning an 'age' to a given category, we experimented with two different schemata. First, we took the median of all active years of people and used that as our age representation. Second, we took the average of all active years of people within a given topic and assigned a one-date age representation for each topic in our dataset, see 3.3.

Figure 3 shows a couple of examples for Physics vs. Chemistry as two distinct categories. Figure 3a and 3b illustrate an interesting behavior that emerges when comparing the branches of the two selected topics in respect to their spread over time. Each node represents a unique branch where the size of the node is amplified by the number of people associated with that branch. It is worth noting that in order for a branch to be considered, a minimum of 5 people per branch was required. Intuitively, we think of Chemistry as a younger subject than Physics. Most recent advances of Chemistry happened in the last decade whereas Physics dates far back in time to some early advances with the rise of several ancient civilizations (similar to mathematics). That intuition was correctly manifested in our results as shown in Figure 3.

Moreover, Figure 3c and 3d show the underlying distribution of people over all branches collectively for a given

(a) (b) (c) (d)

Figure 3: (a) and (b) show each unique branch of a given topic as a node where the size of the node represents the number of people associated with that branch and the y-axis corresponds to the branch's age. On the other hand, (c) and (d) show the underlying people distribution for all branches collectively for a given topic. Each node represents an individual plotted on a timeline using the mean of all years found for that given person.

topic. Each node represents an individual plotted on a timeline using the mean of all years found for that given person. As you can see, Physics had many early breakthroughs with unbelievably smart characters coming out of the dark ages introducing new theories/hypotheses about the universe, whereas Chemistry started to shine with the rise of the technological revolution around the 19th and 20th centuries. These plots were also initially designed as interactive plots, complete with hover effects, such that someone could easily explore the different people and branches. Unfortunately, we realized somewhat late that interactive plots that can only be hosted locally do not lend themselves especially well towards papers or presentations.

3.3. One Number Solution

Lastly, and certainly not 'leastly', we crunched the collective numbers for a topics age down to just one date 3.1. This number, dangerously, represents the mean of all of the people's means, across all of the people found on all branches. Dates across a number of example subjects can be seen in Table 1. This is, of course, in the vain attempt of converting something as ethereal as Wikipedia topics age to our more familiar one number system. Alternatively, as cemented in the 22nd Annual Leadership Convention for Partially Existent Wikipedia Entities (LCPEWE), a (Topic $\vee$ Category)'s age falls under the previously defined regulations for stub articles [1]. Namely, age- when defined as a single number- must represent a randomly generated integer value between 1 and 1000 (inclusive) [3].

4. Discussion

Although our plots and our data may (or may not) reflect a fully detailed and realistic picture of different academic disciplines in the real world, one can arguably say that our results give a rough intuition of their relative presence across time. These findings are, due to the severe lim-

itations of Wikipedia data, only relevant really within the scope of Wikipedia. That being said, it appears that most topics when viewed from the high level of our approach tend to conform to our preexisting intuitions. For example, Geology and Agriculture are relatively old in comparison to Computer Science and Chemistry (which seems reasonable, right?)

Ultimately, most of our efforts were consumed by the scraping process and the various sub problems introduced along the way. In this sense, the more difficult question (which questions to ask) remained unconsidered. Our biggest difficulty therefore was in refining our project down to these explicit questions, and in not getting distracted by the allures of scraping. In future work, if possible, we would like to focus more efforts towards devising new questions to ask about our collected dataset. Specifically, we would like to explore in more depth questions such as, is there a relationship between branch network structure and age, can we combine multiple sources of a topics age in a way that makes sense, and People-Net.

Acknowledgement

Our study came from the principles of the University of Vermont, taught by Professor Peter Dods. We are grateful for the tremendous amount of gratitude we have associated with the mysterious teaching of Davidic Power. Apart from this, the project will not be able to work with Johnny, who has been conspired to transit the transist. No trolls, no computers, no computers, no keyboard ... without a keyboard, how can I write this? The man's heart plans to get the documents at the Steering Committee. Explanation: We would like to present this to the intimidating but revolutionary iCOW Board. In addition, when they receive a positive response, our handwritten publication is published in the popular book.

References

[1] *Proceedings on the more obtuse nature of reality*, volume 22 of *Annual Leadership Convention for Partially Existent Wikipedia Entities*, Wiki-town 2, Wiki-District 8, 842. Springer.

[2] W. contributors. Wikipedia, the free encyclopedia, 2004.

[3] Council of Wiki Leadership. Council regulation (Wiki) no 269/264, 264.