Identifying Birth and Death Years of Authors of Undated Documents using Citations and various Constraints

Dror Mughaz[1,2], Yaakov HaCohen-Kerner[2], Dov Gabbay[1]

[1] Dept. of Computer Science, Bar-Ilan University, 52900 Ramat-Gan, Israel

[2] Dept. of Computer Science, Jerusalem College of Technology, 91160 Jerusalem, Israel

myghaz@yahoo.com, kerner@jct.ac.il, dov.gabbay@kcl.ac.uk

Abstract

There are many implicit details which can be extracted from a text, one of these details is the identification of the era in which the book/paper was written or author of the given document(s) lived.

In this work we will formulate several rules some of them various kinds of "iron-clad", heuristic and sum are greedy constraints defining the birth and death years of an author the rules based on citations in the documents

This problem is important for rabbinic responsa written in Hebrew-Aramaic, which are almost without exception undated and do not contain any bibliographic section, and mentions historical events in the documents are very rare.

Aim

The aim of this novel research is to find in which years an author was born and died, based on his documents and the documents of other authors (whose birth and death years are known) who refer to the author under discussion or are mentioned by him.

Background

Finding in which years an author was born and died can help determine the time frame in which certain documents were written identify an anonymous author and find text that was rewrite/edited.

Citations are a defining feature of many kinds of documents, e.g., academic, legal and religious. Citations included in documents are important information resources of interest to researchers.

Garfield (1965) was the first to propose automatic production of citation indexes, extraction and analysis of citations from corpora of academic papers.

Teufel et al. (2006) use extracted citations and their context for automatic classification of citations to their citation function, the author's reason for citing a given paper.

Methods

We formulated some citation-based constraints to estimate the birth and death years of an author based on undated citations of other authors (whose birth and death years are known) who refer to him or mentioned by him.

A first kind of classification of the constraints:

1. **I** - "Iron-clad" constraints (absolutely true).
2. **H**- Heuristic constraints (almost always true).
3. **G**- Greedy constraints (rather reasonable).

A second kind of classification of the constraints :

1. General citations without cue words.
2. Citing years.
3. Citations with cue words, such as: father, son, rabbi, teacher, student, friend, and "late" ("of blessed memory").

result	Friend	citation
1781	-----	1781
1920	1920	1915
1863	1863	1880

Example of citation with cue word "friend"

A third kind of classification of the constraints:

1 Constraints referring to living authors.

2 Constraints referring to dead authors. In contrast to academic papers, responsa include much more citations to dead authors than to living authors.

Notions and Constants:

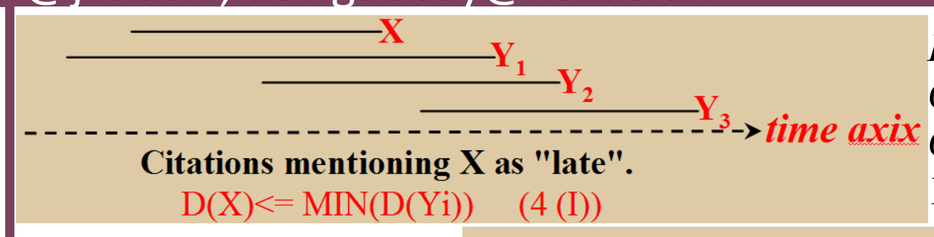
X – The author under consideration **B** – Birth year

Yi – Other authors **D** – Death year

B(X)>= MAX(D(Yi)) (10 (G))

Example of a Greedy constraint

Greedy constraint for defining the birth year based only on authors who were cited by X

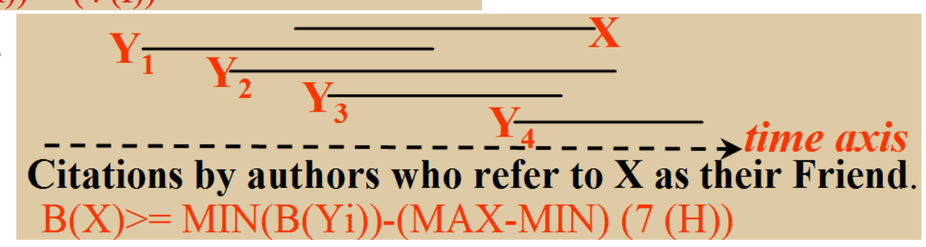


Example of an "Iron-Clad" Constraint

Constraint based on authors Yi that cite X as "late"

Example of a Heuristic-Constraint

Constraint based on authors Yi that cite X



For the Iron, Heuristic and Greedy constraint we added some tunings like the age must be **30<age<100** and **"Current Year".**

The Main Steps of the Model

1. **Cleaning the texts.**
2. **Normalizing the citations in the texts.**
3. **Building indexes.**
4. **Citation identification.**
5. **Performing various combinations of "iron-clad" and heuristic constraints** on the one hand, **and greedy constraints** on the other hand, **to estimate** the birth and death years for each tested author.
6. **adding tunings.**
7. **Calculating averages** for the "iron-clad" and heuristic versions and the greedy versions.

Results - Preliminary Findings

We experiment on four sets of authors:

- 12 scholars containing 10512 files on the average 876 files for each scholar cross 133 years.
- 24 scholars containing 15450 files on the average 643 files for each scholar cross 227 years.

We ran four combinations : (1) Without constant and without years; (2) With constant and without years; (3) Without constant and with years and (4) With constant and with years. Each of the four options above was run on two algorithms iron+heuristic and greedy. For each of them we evaluated the birth years and death the years.

Due to lack of space we present only two tables that contain the greatest number of occurrences of best results.

		Birth				**Death**			
		with constant and without years				without constant and with years			
	# of authors	No refinement	Late	Master	Friend	No refinement	Late	Master	Friend
Iron + Heuristic	12	Age	Age	Age	Age	Age	Age	Age	Age
		24.33	47.8	27.17	**12.79**	**9.67**	16.81	17.33	18.25
	24	Age	Age	Age	Age	no tuning	Age	no tuning	Age
		30.42	40.9	33.35	**22.17**	**13.08**	15.04	23.63	22.73
Greedy	12	const	Age	const	Age	Age	Age	Age	Age
		16.83	22.3	13.67	**13.42**	**10.5**	17.38	19.04	12.38
	24	Age	Age	Age	Age	Age	Age	Age	Age
		14.31	18.3	15.94	**14.13**	23.29	**20.83**	28.46	25.92

Discussion

The use of **"Age"** manipulation usually gives the best results because that where there is an anomaly there is an error and "Age" manipulation reduces it. Due to that author writ quite close to the year of his death the use of Iron+Heuristic algorithm is assessing a good **death year**.

The "friend" refinement for **birth year** estimation gives the best results because friends are more or less the same age, thus they born, roughly, on the same year. Therefore, an author referring to another as his friend it's a good hint for his birth year.

Literature Cited

1) Garfield, E.: Can Citation Indexing be Automated? In: Stevens, M. (ed.) Statistical Association Methods for Mechanical Documentation, Symposium Proceedings, vol. 269, pp. 189–142. National Bureau of Standards Miscellaneous Publication (1965).

2) Teufel, S., Siddharthan, A., Tidhar, D.: Automatic Classification of Citation Function. In: The 2006 Conference on Empirical Methods in Natural Language Processing, ACL, pp. 103–110 (2006).