ISSN 1870-4069

Citation-based Prediction of Birth and Death Years of Authors

Dror Mughaz[1,2], Yaakov HaCohen-Kerner[2], Dov Gabbay[1,3]

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

[2] Lev Academic Center,
Dept. of Computer Science,
Israel

[3] Kings College London,
Dep. of Informatics,
United Kingdom

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

Abstract. This paper presents an unusual approach in text mining and feature extraction for identifying the era of anonymous texts that can help in the examination of forged documents or extracting the time-frame of which an author lived. The work and the experiments concern rabbinic documents written in Hebrew, Aramaic and Yiddish texts. The documents are undated and do not contain any bibliographic sections, which leaves us with an interesting challenge. This study proposes a few algorithms based on keyphrases that enable prediction of the time-frame of which the authors lived based on the temporal organization of references using linguistic patterns. Based on the keyphrases and the citations we formulated various types of "Iron-clad", Heuristic and Greedy constraints defining the birth and death years of an author that lead to an interesting classification task. The experiments that were applied on corpora containing texts authored by 12, 24 and 36 rabbinic authors show promising results.

Keywords: Citation analysis, text mining, Hebrew-Aramaic documents, knowledge discovery, time analysis, undated citations, undated documents.

1 Introduction

Extracting the era of a book/manuscript and determining who the author who wrote it, is very challenging and essential problems, the use of citations can help us to do that. Citations provide a lot of important information to studies in different areas such as academic, legal and religious. Therefore, extracting and analyzing citations

automatically is growing and gaining momentum. Search engines and digitized corpora enable identification and extraction of citations. Hence, citation analysis has an increased importance.

The use of citations is not limited to academic papers it is used also in rabbinic responsa (questions and detailed rabbinic answers). The citations that includes in rabbinic texts are more hard to define and to pull out than citations in academic papers because: (1) There is no bibliography at the end of a responsa; (2) The complex morphology of Hebrew, Aramaic and Yiddish; (3) NLP in Hebrew Aramaic and Yiddish has been little studied; (4) Many citations in Hebrew-Aramaic-Yiddish documents are ambiguous; (5) these citations are undated and (6) Plenty of different styles and syntactic used to display citations [1].

The Semitic languages are substantially different from the Latin languages. Due to that, Hebrew, which is Semitic language, will have very different processing from English in a variety of linguistic aspects. In Hebrew the writing direction is from right-to-left which is different from the Latin languages. Also, in Hebrew there are no vowels thus there are lots of ambiguous words. For example, the word דלק can be read as Delek, which means a fuel or Dalak, which means burned and as well to-chase; another example is the word ספר, which can be can be read as Safar, which means counted, or Sapar, which means barber, or Sefer which means a book [2].

The morphology of Hebrew, compared to the morphology of English, is plentiful. Most the function words in English are affixes in Hebrew. The Hebrew has a normal form, which called a "root" which has many different forms that sometimes not just add characters to it but also change the structure of the root pattern and the meaning of the word. (For example, לשמש = ל+שמש (La-Shemesh, which means to the sun (NP)) or לשמש (Leshamesh, which means to serve (adv))) [2].

There is relatively a high rate of acronyms and abbreviations in Hebrew texts. The number of abbreviations is about 17,000, compared to 40,000 lexical entries in Hebrew [3] it is relatively high, hence, another type of ambiguity. In Hebrew Rabbinic texts the use of acronyms and abbreviations are by far more plentiful than regular Hebrew texts.

HaCohen-Kerner et al. [3, 4, 5, 6] construct a system to disambiguate Hebrew and Aramaic acronyms for classical Jewish texts, mostly in pre-Modern Hebrew. They used an existing dictionary of acronym expansions and with the use of machine learning techniques they achieved high accuracy of automatic acronym expansion.

HaCohen-Kerner et al. [7] showed that manual disambiguation of an acronym is greatly time-consuming and even for highly trained human experts it's a challenging task.

Liebeskind et al., [8; 9;10] addressed the task of thesaurus construction, aiming to enable potential users of the Responsa Project to search for modern terms and obtain semantically related terms from earlier periods in history.

They proposed an algorithmic scheme for generating a co-occurrence based thesaurus in Morphologically Rich Languages (MRL) and demonstrate its empirical benefit for the Hebrew diachronic thesaurus [10]. Then, they introduced a semi-automatic iterative Query Expansion (QE) scheme that increases recall and the effectiveness of lexicographer manual effort [8]. Later, their scheme was extended to deal with Multi-Word Expressions [9].

This research uses undated citations of other dated authors in order to assess the date of undated documents. The assessment based on a several rules of different level of

certainty: "iron-clad", heuristic and greedy. The rules are based on citations: (I) typical citations with no cue words and (II) citations with cue words such as: "late" ("of blessed memory"), "friend" and "rabbi".

Previous studies in this area [11, 12, 13] were made on: 12 or 24 authors, on two different years' intervals and without normalizing the difference between those intervals. We expend these studies from several aspects: using a much larger corpus, increasing the number of authors, normalizing the amplitudes years (to have a comparison) and examine the constants that are part of the heuristic rules.

This paper is organized as follows: Section 3 presents various constraints of different degree of certainty: "iron-clad", heuristic and greedy constraints that are used to estimate the birth and death years of authors. Section 4 describes the model. Section 5 introduces the tested dataset, the results of the experiments and their analysis. Section 6 summarizes, concludes and proposes future directions.

2 Related Works

The first to propose extraction of citation automatically for indexing and analysis from academic corpora papers was Garfield [14]. Berkowitz and Elkhadiri [15] pull out titles and author names from articles. Giuffrida et al. [16] extract author names from metadata of computer science articles by knowledge-based framework. Seymore et al. [17] extract author name by hidden Markov models from a small corpus of computer science articles.

Teufel et al. [18] classify citations to their function by extraction automatically the citations and their context (the reason for citing the paper). Tan et al. [19] disambiguate author of the results of automatically-crafted web searches. Improvement of extracting terms using citations has been done. Bradshaw [20] uses a fixed window round citations to extract terms. Ritchie et al. [21] improve retrieval effectiveness by selecting text from around citations in order to extract good index terms.

Temporal citation-related problems were done on the traditional Western scientific literature. Popescul et al. [22] present an approach for identify and cluster temporal deal with hyper-linked scientific texts databases. Kolomiyets et al. [23] and Bethard et al. [24] dealt with the issue of creating a timeline, the timeline is for each text on its own and relating to that text they create chronological order. Schwartz et al. [25] and Wen et al. [26] studied texts on social networks addressing a story/documentation of the specific process (medical, etc.) that occurs to individual.

Many studies in information retrieval have been done on citations (e.g. IR [27; 28; 21; 29; 30; 15; 2]). However, our study we are addressing much harder issue of citations that included in rabbinic texts. Mughaz et al. [13] present's approach of cross generation and citation-based method to date undated authors. Their experiment was based on a small corpus of only 12 authors and 24 authors.

3 Citation-Based Constraints

In this part we show the citation-based formulated rules in order to estimate the birth and death years of a writer X (the outcomes point to particular years) based on his texts

and on different writers' (Yi) texts who mention X or one of his books. We are assuming that the death years and birth years of all writers (Yi) are known, but those of the examined writer (X). Now we are giving constants and notions: X – The author under examination, Yi – Other writers, B – Birth year, D – Death year, MIN – Minimal age (currently 30) of a rabbinic writer when he starts to write his response, MAX – Maximal life period (currently 100) of a rabbinic author and RABBI_DIS – The age gap between rabbi and his scholarly student (currently 20).

The assessment of MIN, MAX, RABBI_DIS constants are heuristic but they are reasonable to a typical responsa authors' way of living. There are several types of citations: general citations without any cue word and citations with cue words, as: "Rabbi", "Friend", and "Late" ("of blessed memory"). The citations are divided to two kinds: those who citing living authors and those who citing dead authors. Unlike academic papers, the responsa contain much more citations to dead authors than to living authors.

We will introduce citation-based constraints of different degrees of certainty: "iron-clad" (I), heuristic (H) and greedy (G). "Iron-clad" constraints are absolutely true, without any exception. Heuristic constraints are almost always true. Exceptions can occur when the heuristic estimates for MIN, MAX and RABBI_DIS are incorrect. Greedy constraints are rather reasonable constraints for responsa authors. However, sometimes wrong estimates can be drawn while using these constraints. Each constraint will be numbered and its degree of certainty will be presented in brackets.

3.1 "Iron-Clad" and Heuristic Constraints

First of all, we present two general heuristic constraints based on authors that cite X, which are based on regular citations (i.e., without mentioning special cue words, e.g., "friend" and "rabbi").

General constraint based on authors that were cited by X:

$$D(X) >= MAX(B(Yi)) + MIN \quad (1\ (H))$$

X must be alive when he cited Yi, so we can use the earliest possible age of publishing of the latest born author Y as a lower estimate for X's death year.

General constraint based on authors that cite X:

$$B(X) <= MIN(D(Yi)) - MIN \quad (2\ (H))$$

All Yi must have been alive when they cited X, and X must have been old enough to publish. Therefore, we can use the earliest death year amongst such authors Yi as an upper estimate of X's earliest possible publication age (and thus his birth year).

General constraints based on references to year Y that were cited by X, later we will address it as a "years-feature":

$$D(X) >= MAX(Y) + MIN\ (3\ (H))$$

X must be alive when he mentioned the year Y, we can use the most recent year referred by X to evaluate the death year of X as estimation for X's death year.

Because writer, in a lot of cases dos not write until his “last day” we add heuristic constant “MIN”.

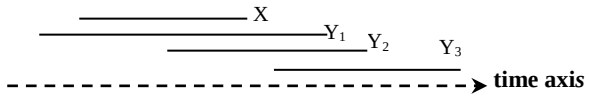

Fig. 1. Citations mentioning X as "late".

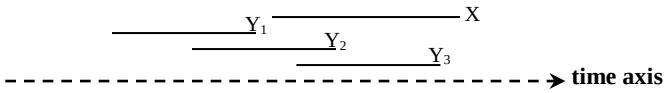

Fig. 2. Citations by X who mentions others as "late".

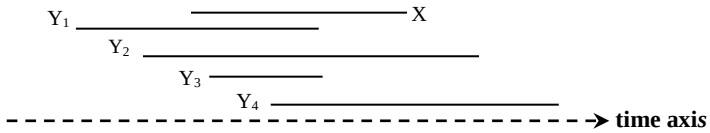

Fig. 3. Citations by authors who refer to X as their Friend/Rabbi.

Posthumous citation constraints

Posthumous constraints estimate the birth and death years of an author X based on citations of authors who refer to X as "late" ("of blessed memory") or on citations of X who mentions other authors as "late". Fig. 1 describes possible situations where various kinds of authors Yi (i=1, 2, 3) refer to X as "late". The lines depict authors' life spans where the left edges represent the birth years and the right edges represent death years. In this case (as all Yi refer to X as "late"), we know that all Yi died after X, but we do not know when they were born in relation to X's birth. Y1 was born before X's birth; Y2 was born after X's birth but before X's death; and Y3 was born after X's death:

$$D(X) <= MIN(D(Yi)) \quad (4\ (I))$$

However, we know that X must have been dead when Yi cited him as "late", so we can use the earliest born such Y's death year as an upper estimate for X's death year. Like all authors, dead authors of course have to comply to constraint (2) as well.

Now look at the cases where the author X, we are studying refers to other authors Yi as "late". Fig 2 describes possible situations where X refers to various kinds of authors Yi (i = 1, 2, 3) as "late". All Yi died before X's death (or maybe X is still alive). Y1 died before X's birth; Y2 was born before X's birth and died when X was still alive; and Y3 was born after X's birth and died when X was still alive:

$$D(X) >= MAX(D(Yi)) \quad (5\ (I))$$

X must be alive after the death of all Yi who were cited as "late" by him. Therefore, we can use the death year of the latest-born such Y as a lower estimate for X's death year.

$$B(X) >= MAX(D(Yi)) - MAX \quad (6\ (H))$$

X may be born after the death year of the latest-dying person, who X wrote about. Thus, we use the death year of the latest-born such Y minus his max life-period as a lower estimate for X's born year.

Contemporary citation constraints

Contemporary citation constraints calculate the upper and lower bounds of the birth year of an author X based only on citations of known authors who refer to X as their friend/rabbi. This means there must have been at least some period in time when both were alive. Fig 3 describes possible situations where various kinds of authors Yi refer to X as their friend/rabbi. Y1 was born before X's birth and died before X's death; Y2 was born before X's birth and died after X's death; Y3 was born after X's birth and died before X's death; and Y4 was born after X's birth and died after X's death. Like all authors, contemporary authors of course have to comply to constraints 1 and 2 as well:

$$B(X) >= MIN(B(Yi)) - (MAX - MIN) \quad (7\ (H))$$

All Yi must have been alive when X was alive, and all of them must have been old enough to publish. Therefore, X could not be born MAX-MIN years before the earliest birth year amongst all authors Yi:

$$D(X) <= MAX(D(Yi)) + (MAX - MIN) \quad (8\ (H))$$

Again, all Yi must have been alive when X was alive, and all of them must have been old enough to publish. Thus, X could not be alive MAX-MIN years after the latest death year amongst all authors Yi.

3.2 Greedy Constraints

Greedy constraints bounds are sensible in many cases, but can sometimes lead to wrong estimates.

Greedy constraint based on authors who are mentioned by X:

$$B(X) >= MAX(B(Yi)) - MIN \quad (9\ (G))$$

Many of the citations in our research domain relate to dead authors. Thus, most of the citations mentioned by X relate to dead authors. That is, many of Yi were born before X's birth and died before X's death. Therefore, a greedy assumption will be that X was born no earlier than the birth of latest author mentioned by X; but because that may be at least one case where Y was born after that X was born so we subtract MIN.

Greedy constraint based on references to year Y that were cited by X, later we will address it as a "years-feature":

$$B(X) >= MAX(Y) - MINy \quad (10\ (G))$$

X reminds years he usually writes the current year in which he wrote the document or several years before. Most of the time the maximum year, Y, reduces MIN is larger than X's born year because of that the “MIN” in (10 (G)) cannot be as the “MIN” in the other constraint so is “MINy” (currently 60).

Greedy constraint based on authors who refer to X:

$$D(X) <= MIN(D(Yi)) - MIN \quad (11\ (G))$$

As mentioned above, most of the citations mentioned by Yi relate to X as dead. Therefore, most of Yi die after X's death. Therefore, a greedy assumption will be that X died no later than the death of the earliest author who refers to X minus MIN.

Constraints refinements 9-11 are presented by constraints 12-17. Constraints 12-14 are due to X citing Yi and Constraints 15-17 are due to Yi citing X.

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

$$B(X) >= MAX(D(Yi)) - MIN \quad (12\ (G))$$

When taking into account only citations that are cited by X, most of the citations, relate to dead authors. That is, most of Yi died before X's birth; in addition, an author doesn't write from his birth but usually until near his death. Therefore, a greedy assumption will be that X was born no earlier than the death of the latest author mentioned by X minus MIN.

Greedy constraint for defining the birth year based only on authors who are mentioned by X as a "friend":

$$B(X) <= MIN(B(Yi)) + RABBI\_DIS \quad (13\ (G))$$

When taking into account only citations that are mentioned by X, which relate to contemporary authors, a greedy constraint can be that X was born no later than the birth of the earliest author mentioned by X as a friend. Because many times older author is mentioning young author as a friend we need to add RABBI_DIS.

Greedy constraint for defining the birth year based only on authors who are mentioned by X as a "rabbi":

$$B(X) <= MIN(B(Yi)) + RABBI\_DIS \quad (14\ (G))$$

When taking into account only citations that are mentioned by X, which relate to contemporary authors, a greedy constraint can be that X was born not later than the birth of the earliest author mentioned by X as a RABBI. Because of the age difference between the rabbi and his student is about 20 years we need to add RABBI_DIS.

Greedy constraint for defining the death year of X based only on authors who cited X as "late":

$$D(X) <= MIN(B(Yi)) + MIN \quad (15\ (G))$$

When taking into account only citations that are mentioned by Yi who relate to X as "late", a greedy assumption can be that X died no later than the birth of the earliest author who cited X as "late" and because author doesn't writes from his birth we need to add MIN.

Greedy constraint for defining the death year of X based only on authors who cited X as a "friend":

$$D(X) >= MAX(D(Yi)) - RABBI\_DIS \quad (16\ (G))$$

When taking into account only citations that are mentioned by Yi who cited X as a friend, all Yi must have been alive when X was alive, and all of them must have been old enough to publish and many times older author is mentioning young author as a friend but the opposite never happen. Therefore, a greedy assumption will be that X

died no earlier than the death of the latest author who cited X as a friend minus RABBI_DIS.

Greedy constraint for defining the death year of X based only on authors who cited X as a "rabbi":

$$D(X) >= MAX(D(Yi)) - RABBI\_DIS \quad (17\ (G))$$

It is the same principle as the constraint for defining the born year but because hear the student mention the rabbi we need to reduce RABBI_DIS.

3.3 Birth and Death Year Tuning

Application of the Heuristic and Greedy constraints can lead to anomalies, such as an author's decease age being unreasonably old or young. Another possible anomaly is that the algorithm may yield a death year greater than the current year (i.e. 2014). Therefore, we added some tuning rules: D – death year, B – born year, age = D-B.

Current Year: if (D > 2016) {D = 2014}, i.e., if the current year is 2014 the algorithm must not give a death year greater of 2014.

Age: if (age > 100) {z = age – 100; D = D – z/2; B = B + z/2} if (age < 30) {z = 30 – age; D = D + z/2; B = B - z/2}. Our assumption is that an author lived at least 30 years and no more than 100 years. Thus, if the age according to the algorithm is greater than 100, we take the difference between that age and 100, then we divide that difference by 2 and normalize D and B to result with an age of 100.

4 The Model

The main steps of the model are presented below. Most of these steps were processed automatically, except for steps 2 and 3 that were processed semi-automatically.

1 **Cleaning the texts**. Since the responsa may have undergone some editing, we must make sure to ignore possible effects of differences in the texts resulting from variant editing practices. Therefore, we eliminate all orthographic variations.

2 **Normalizing the citations in the texts**. For each author, we normalize all kinds of citations that refer to him (e.g., various variants and spellings of his name, books, documents and their nicknames and abbreviations). For each author, we collect all citation syntactic styles referred to him and then replace them to a unique string.

3 **Building indexes**, e.g., authors, citations to "late"/friend/rabbis and calculating the frequencies of each item.

4 **Citation identification** into various categories of citations, including self- citations.

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 **Calculating averages** for the best "iron-clad" and heuristic version and the best greedy version**.**

Table 1. Birth average distance. without years-feature.

	# of authors	No refinement	Late	Rabbi	Friend
Iron + Heuristic	12	Age	const	Age	Age
		46.56	60.97	37.5	**26.32**
	24	Age	const	Age	Age
		34.93	33.65	28.58	**26.13**
	36	Age	const	Age	Age
		31.47	27.63	23.39	**19.6**
Greedy	12	Age	Age	const	Age
		30.22	39.85	**25.28**	28.37
	24	Age	Age	Age	Age
		15.28	19.98	17.46	15.71
	36	Age	Age	Age	Age
		12.46	20.51	14.24	14.4

Table 2. Death average distance. without years-feature.

	# of authors	No refinement	Late	Rabbi	Friend
Iron + Heuristic	12	Age	no tuning	const	Age
		43.1	41.5	45.1	**28.3**
	24	Age	no tuning	const	Age
		23.98	**19.9**	32.7	24
	36	Age	no tuning	Age	Age
		17.72	18.2	20	17
Greedy	12	Age	Age	Age	Age
		30.53	31.8	38.2	**23.9**
	24	Age	Age	Age	Age
		30.68	**22.9**	32.1	24.9
	36	Age	Age	Age	Age
		19.68	22.3	20.4	**17.8**

5 Experimental Results

The documents of the examined corpus were downloaded from the Bar-Ilan University's Responsa Project[1]. The examined corpora contain 24,111 responsa written by 36 scholars, averaging 670 files for each scholar. These authors lived over a period of 250 years (1765–2015). These files contain citations; each citation pattern can be expanded into many other specific citations [12].

The citation recognition in this research is done by comparing each word to a list of 339 known authors and many of their books. This list of 25,801 specific citations that relate to names, nick names and abbreviations of these authors and their writings. Basic citations were collected and all other citations were produced from them.

We divide the data into three sets of authors documents (1) 12 scholars: containing 10,561 files, 880 files on average for each scholar spread over 135 years (1880–2015); (2) 24 scholars: containing 15,495 files, 646 files on average for each scholar spread over 229 years (1786–2015) (the set of 24 authors contains the set of 12 authors); and (3) 36 scholars: containing 24,111 files, 670 files on average for each scholar spread over 250 years (1765–2015) (the set of 36 authors contains the set of 24 authors). The research question (prediction of birth and death years of authors based on undated citations) that we face is relatively novel.

[1] The Global Jewish Database (The Responsa Project at Bar-Ilan University). Http://www.biu.ac.il/ICJI/Responsa.

Thus, there is no basis for comparison to assess the accuracy of the results. Because of this we use the distance function, i.e., we will measure the distance between the real birth and death years of the authors and the assessments of the algorithms in order to evaluate the results.

Because we have three groups of authors; each one with a different span of years, as mentioned before, we have to normalize the results in order to compare between them. The results of the original 12 authors are multiplied by 1.85 when compared to the results of 36 authors (the amplitude years of 12 authors is 135, the amplitude years of 36 authors is 250, result 250 / 135 ~ 1.85), we did the same for the set 24 authors (the amplitude years of 24 authors is 229, the amplitude years of 36 authors is 250, result 250 / 229 ~ 1.09).

The results that appear in the following tables, each table shows results of two algorithms - Iron+Heuristic (sub-section 3.1) and Greedy (sub-section 3.2). Each algorithm was performed on three groups of authors: 12 authors, 24 authors, and 36 authors. For both algorithm executions there are results containing estimated years of birth and death. The results shown in the tables are the best birth/death date deviation results. In every quarter table there are four columns: a deviation without refinement, a deviation with the "Late" refinement, with the "Rabbi" refinement, and with the "Friend" refinement (Section 3).

In addition, we used two manipulations - Age and Current year (sub-section 3.3). The bold cells contain the best results. The following four tables contain the results of the evaluations of birth and death years of the two algorithms, i.e., the Greedy algorithm and the Iron+Heuristic algorithm, with the use of the years-feature (see (3 (H)) and (10(G))), total of 96 results.

The Age manipulation gives the best results with 76% for all refinements, in both algorithms, with or without constants, i.e. in the two tables (73/96=0.76). It doesn't necessarily mean that to all the authors Age manipulation was done, but for some of them it was necessary.

This manipulation is effective, because it is only used when estimate is erratic; therefore, a manipulation is necessary and usually it improves the results. For example, in the Greedy algorithm, when using years with the "Rabbi" refinement for 36 authors, the estimated birth years and death years are improved in 44 results out of 72 (29 for birth year estimates and 15 for death year estimates); the average improvement estimate was 15.9 years for all 72 results (36 birth years and 36 death years).

In the Iron+Heuristic algorithm, when using the years and "friend" refinements for 36 authors, the estimated birth years and death years are improved in 45 results out of 72 (22 for birth year estimates and 23 for death year estimates); the average improvement estimate was 11.13 years for all 72 results (36 birth years and 36 death years). To summarize, the Age manipulation is very helpful for the birth and death year reckoning.

First, we will analyze the algorithms from a general perspective, consistency; there are two aspects in terms of consistency: (1) at the level of the best results of the Iron+ Heuristic/Greedy algorithms for birth/death year assessments and (2) at the level of a specific refinement. (1) Are the best results of the 36 authors are better than the best results of the 24 authors and also are the best results of the 24 authors are better than the best results of the 12 authors? (2) For each specific refinement, is the result of the

Table 3. Birth average distance.

	# of authors	No refinement	Late	Rabbi	Friend
Iron + Heuristic	12	Age	const	Age	Age
		63.52	58.28	60.78	**26.32**
	24	Age	const	Age	Age
		42.96	31.72	33.64	**27.47**
	36	Age	const	Age	Age
		38.67	26.1	30.89	**21.18**
Greedy	12	Age	const	const	Age
		23.67	24.36	**22.82**	28.67
	24	const	Age	const	Age
		12.67	**12.38**	15.44	16.74
	36	const	Age	Age	Age
		11.63	**11.35**	14.17	15.35

with years-feature.

Table 4. Death average distance.

	# of authors	No refinement	Late	Rabbi	Friend
Iron + Heuristic	12	Age	no tuning	const	Age
		18.35	22.5	34.3	28.3
	24	no tuning	no tuning	Age	Age
		14.4	20.1	33.1	25.3
	36	no tuning	no tuning	Age	Age
		12.86	18.2	20.2	17.7
Greedy	12	Age	Age	Age	Age
		19.89	32.6	35.8	23.9
	24	Age	Age	Age	Age
		25.55	**22.9**	31.2	27.2
	36	Age	Age	Age	Age
		23.44	**21**	28.7	25

with years-feature.

36 authors better than the result of the 24 authors and for the same refinement, is the result of the 24 authors is better than result of the 12 authors.

In the Iron+Heuristic algorithm there is an almost complete consistency in the two aspects, namely the more authors and more information the deviation results become better. For example, from the first perspective, in table 2 the best result of 36 authors is 17.72, the best result of 24 authors is 19.92, and the best result of 12 authors is 28.33; the more authors the better the predictions. Example for the second perspective, in table 2 with the Rabbi refinement, the best result of 36 authors is 20.03, the best result of 24 authors is 32.73, and the best result of 12 authors is 45.13.

There is one case of inconsistency at the Iron+Heuristic algorithm in table 3 with the friend refinement, where the result became worse but with a relatively small deviation (about 1.1). However, compared with the Iron+Heuristic algorithm, in the Greedy algorithm there is greater inconsistency in both levels. For instance, from the first aspect, table 4 shows that the best result of 36 authors is 20.98, the best result of the 24 authors is 22.87, and the best result of 12 authors is 19.8; from the second aspect, reviewing table 4 in the Rabbi refinement, the best result of 36 authors is 24.98, the best result of 24 authors is 27.23, and the best result of 12 authors is 23.9.

A possible explanation is that the Greedy algorithm is an intuitive therefore its consistency is "not his forte"; In contrast, the Iron+Heuristic algorithm is mathematically "committed" (up to the heuristics) and therefore it is much more consistent. In conclusion, the more information the Iron+Heuristic algorithm is more stable and more consistent.

Table 5. Birth average distance without years, Mughaz et al. [13].

	# of authors	No refinement	Late	Rabbi	Friend
Iron + Heuristic	12	49.29	93.43	51.15	23.79
	24	37.68	44.8	35.27	24.94
Greedy	12	31.01	54.33	39.38	33.09
	24	24.45	30.25	26.74	22.32

Table 6. Death average distance without years, Mughaz et al. [13].

	# of authors	No refinement	Late	Rabbi	Friend
Iron + Heuristic	12	45.25	31.19	45.57	33.87
	24	26.4	16.54	28.53	23.14
Greedy	12	29.91	32.01	38.28	23.03
	24	30.78	22.91	32.15	26.76

Table 7. Birth average distance without years, Current results.

	# of authors	No refinement	Late	Rabbi	Friend
Iron + Heuristic	12	46.56	60.97	37.5	26.32
	24	34.93	33.65	28.58	26.13
Greedy	12	30.22	39.85	25.28	28.37
	24	15.28	19.98	17.46	15.71

Table 8. Death average distance without years, Current results.

	# of authors	No refinement	Late	Rabbi	Friend
Iron + Heuristic	12	43.1	41.47	45.13	28.33
	24	23.98	19.92	32.73	23.98
Greedy	12	30.53	31.84	38.16	23.9
	24	30.68	22.87	32.09	24.89

When we use the Iron+Heuristic algorithm in order to evaluate the birth years we can see from tables 1 and 3 that unequivocally that the friend refinement always gives the best results compared to the other refinements whether we use the years-feature or whether we do not. It is arises from formula (7(H)), that the output of the core of formula (7(H)), i.e., MIN(B(Yi)), is much closer to the real birth year so that the use of a constant in (7(H)), i.e., (MAX-MIN), will lead to better results relative to the other refinements.

For example, in formula (7(H)), (which serves the Rabbi refinement and the friend refinement) the MIN(B(Yi)) of the Rabbi refinement is less than the MIN (B(Yi)) of the friend refinement (usually the Rabbi of X born before the friend of X), therefore, the average results of the Rabbi refinement will not be as good as the results of the friend refinement; similarly in the "late" refinement in formula (6(H)). As opposed to the Iron+Heuristic algorithm, the Greedy algorithm does not have consistency.

In the Greedy algorithm, the best result of the set of 12 authors was obtained using the friend refinement; the best results of sets of 24 and 36 authors, were obtained without the use of years-feature are using “no refinement” and with the use of years-feature are using the "late" refinement.

Although the Iron+Heuristic algorithm is more stable and consistent, the results of the Greedy algorithm are better (both with and without the use of the years-feature). When we analyze the differences between the numerical results of the two algorithms, we see that the results of the Greedy algorithm are better in all cases except two cases

Table 9. Birth average distance with years, Mughaz et al. [13].

	# of authors	No refinement	Late	Rabbi	Friend
Iron + Heuristic	12	94.25	93.5	74.55	23.72
	24	59.04	44.8	45.24	26.79
Greedy	12	95.79	54.33	76.8	33.09
	24	54.36	30.25	42.67	25.85

Table 10. Death average distance with years, Mughaz et al. [13].

	No refinement	Late	Rabbi	Friend
Iron + Heuristic	17.99	31.27	32.23	33.95
	14.39	16.54	25.99	25
Greedy	19.53	32.33	35.41	23.03
	25.62	22.91	31.31	28.51

Table 11. Birth average distance without years, Current results.

	# of authors	No refinement	Late	Rabbi	Friend
Iron + Heuristic	12	63.52	58.28	60.78	26.32
	24	42.96	31.72	33.64	27.47
Greedy	12	23.67	24.36	22.82	28.67
	24	12.67	12.38	15.44	16.74

Table 12. Death average distance without years, Current results.

	No refinement	Late	Rabbi	Friend
Iron + Heuristic	18.35	22.51	34.3	28.33
	14.4	20.05	33.1	25.3
Greedy	19.89	32.61	35.77	23.9
	25.55	22.87	31.22	27.23

of 12 authors with the friend refinement (a difference average of 2.2 years, for the two cases).

The average years' deviation without the use of years-feature for the Greedy algorithm is 23.7 and for the Iron+Heuristic algorithm is 30.34. The average years' deviation with the use of years-feature for the Greedy algorithm is 21.93 and for the Iron+Heuristic algorithm is 30.28.

When we are considering only the best results for each set of authors, without the use of the years-feature in the Iron+Heuristic algorithm the average years' deviation is 23 and in the Greedy algorithm the average years' deviation is 19.59; i.e., the results of the Greedy algorithm are more accurate. When we use the years-feature, the years' deviations of the Iron+Heuristic and the Greedy algorithms are 20.1 and 18.38, respectively. In conclusion, the numerical results of the Greedy algorithm are better but the results of the Iron+Heuristic algorithm are more stable and consistent.

Current research versus previous research

In this research, various additions are presented comparing to Mughaz et al. [13] (We are competing with [13] and not with 12 or 11 because 13 is the advanced among them):

1 There are three corpora of responsa composed by 12, 24 and 36 authors, instead of tow corpora (12 and 24 authors),

2 There is a use of three different time intervals instead of only two (Certainly we normalize the results accordingly),

3 We analyzed the stability and consistency of the two algorithms,

4 We checked and used different values of constants where at Mughaz et al. [13] they use the same values of constants,

5 We extend the formula of "years-feature", i.e., (3 (H)) and (10(G)).

Mughaz et al. [13] examined a corpus, which includes 15,450 responsa where 10,512 responsa of them were written by 12 scholars and 15,450 responsa were written by 24 scholars. In this research, the corpus contains 24,111 responsa where 10,561 responsa of them were written by 12, 15,495 responsa were written by 24 scholars, and 24,111 responsa written by 36 scholars. The 15,450 responsa used in Mughaz et al. [13] are included this research.

When we consider the results of Mughaz et al. [13] study versus the results of the current study we see that we have improved 44 results out of 64 results (of 12 and 24 authors), i.e., an improvement of 69%. In more details: The Iron+Heuristic algorithm presents better results of 5.49 years on average (for all the 36 experiments); with using the years-feature the results are better in 7.39 years on average, without using the years-feature the results are better in 3.6 years on average. In 22 cases out of 36 the results are better and in 14 cases are worse.

The Greedy algorithm presents better results of 10.2 years on average (for all the 36 experiments); with using the years-feature the results are better in 16 years on average, without using the years-feature the results are better in 4.39 years on average. In 30 cases out of 36 the results are better and in 6 cases are worse.

6 Summary, Conclusions and Future Work

We investigate the estimation of the birth and death years of the authors using undated citations referring to them or written by them. This research was performed on a special case of documents (i.e., responsa), where special writing rules are applied. The estimation was based on the author's documents and documents of other authors who refer to the discussed author or are mentioned by him. To do so, we use various kinds of iron-clad, heuristic and greedy constraints.

In this research we show an improvement of 44 results out of 64 results, i.e., an improvement of 69%.The examination of the estimation of the birth and death year indicate that the Greedy algorithm has been obtain better results than of the Iron+Heuristic algorithm but the stability and consistency Iron+Heuristic algorithm is better.

Regarding the estimation of the birth and death years of an author X, it is important to point that citations mentioned by X or referring to X are more suitable to assess the "birth" and "death" writing years of X rather than his real birth and death years.

This model can be applied with suitable changes to similar research problems that might be relevant for some historical document collections.

We plan to improve the assessment of the birth and death years of authors by: (1) Combining and testing new combinations of iron-clad, heuristic and greedy constraints, (2) Improving existing constraints and/or formulating new constraints, (3) Defining and applying heuristic constraints that take into account various details included in the responsa, e.g., events, names of people, concepts, special words and collocations that

can be dated, (4) Conducting additional experiments using many more responsa written by more authors is supposed to improve the estimates, (5) Checking why the iron-clad, heuristic and greedy constraints tend to produce more positive differences, and (6) Testing how much of an improvement we got from a correction of the upper bound of D(x) and how much we will at some point use it for a corpus with long-dead authors.

References

1. HaCohen-Kerner, Y., Schweitzer, N., Mughaz, D.: Automatically identifying citations in hebrew-aramaic documents. Cybernetics and Systems: An International Journal, vol. 42 no. 3, pp. 180–197 (2011) doi: 10.1080/01969722.2011.567893
2. Wintner, S.: Hebrew computational linguistics: Past and future. Artificial Intelligence Review. vol. 21, no. 2, pp. 113–138 (2004) doi:10.1023/B:AIRE.0000020865.73561.bc
3. HaCohen-Kerner, Y., Kass A., Peretz, A.: Baseline methods for automatic disambiguation of abbreviations in Jewish law documents. In: Proceedings of the 4th International Conference on Advances in Natural Language, pp. 58–69 (2004) doi:10.1007/978-3-540-30228-5_6
4. HaCohen-Kerner, Y., Kass A., Peretz, A.: Abbreviation disambiguation: Experiments with various variants of the one sense per discourse hypothesis. In: Proceedings of the Application of Natural Language to Information Systems (NLDB'08), pages 27–39 (2008) doi: 10.1007/978-3-540-69858-6_5
5. HaCohen-Kerner, Y., Kass A., Peretz, A.: Combined one sense disambiguation of abbreviations. ACL, pp. 61–64 (2008)
6. HaCohen-Kerner, Y., Kass A., Peretz, A.: HAADS: A hebrew aramaic abbreviation disambiguation system. Journal of the American Society for Information Science and Technology, vol. 61, no. 9, pp. 1923–1932 (2010)
7. HaCohen-Kerner, Y., Kass A., Peretz, A.: Initialism disambiguation: Man versus machine. Journal of the American Society for Information Science and Technology, vol. 64, no. 10, pp. 2133–2148 (2013) doi: 10.1002/asi.22909
8. Mughaz, D., HaCohen-Kerner, Y., Gabbay, D.: Estimating the birth and death years of authors of undated documents using undated citations. In: International Conference on Natural Language Processing. Springer, pp. 138–149 (2010) doi: 10.1007/978-3-642-14770- 8_17
9. Mughaz, D., HaCohen-Kerner, Y., Gabbay, D.: When text authors lived using undated citations. In: Information Retrieval Facility Conference, Springer, vol. 8849, pp. 82–95 (2014) doi: 10.1007/978-3-319-12979-2_8
10. Mughaz, D., HaCohen-Kerner, Y., Gabbay, D.: Key-Phrases as means to estimate birth and death years of jewish text authors. Semanitic Keyword-based Search on Structured Data Sources, Springer, pp. 108–126 (2015) doi: 10.1007/978-3-319-27932-9_10
11. Garfield, E.: Can citation indexing be automated? Statistical Association Methods for Mechanical Documentation, In: Symposium Proceedings, National Bureau of Standards Miscellaneous Publication, vol. 269, pp. 189–142 (1965)
12. Berkowitz, E., Elkhadiri, M. R.: Creation of a style independent intelligent autonomous citation indexer to support academic research. pp. 68–73 (2004)
13. Giuffrida, G., Shek, E. C., Yang, J.: Knowledge-based metadata extraction from postscript files. In: Proceedings of the 5th ACM conference on Digital libraries, ACM, pp. 77–84 (2000) doi:10.1145/336597.33663
14. Seymore, K., McCallum, A., Rosenfeld, R.: Learning hidden Markov model structure for information extraction. In: Workshop on Machine Learning for Information Extraction, pp. 37–42 (1999)

15. Bradshaw, S.: Reference directed indexing: redeeming relevance for subject search in citation indexes. Research and advanced technology for digital libraries, Springer, pp. 499–510 (2003)
16. Ritchie, A., Teufel, S., Robertson, S.: Using terms from citations for IR: some first results. In: The European Conference for Information Retrieval (ECIR), pp. 211–221 (2007) doi: 10.1007/978-3-540-78646-7_21
17. Ritchie, A., Robertson, S., Teufel, S.: Comparing citation contexts for information retrieval. In: 17th ACM Conference on Information and Knowledge Management (CIKM), pp. 213–222 (2008) doi: 10.1145/1458082.145811
18. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: Conference on Empirical Methods in Natural Language Processing, pp. 103–110 (2006)
19. Kolomiyets, O., Bethard, S., Moens, M. F.: Extracting narrative timelines as temporal dependency structures. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 88–97 (2012)
20. Bethard, S., Kolomiyets, O., Moens, M. F.: Annotating story timelines as temporal dependency structures. In: Proceedings of the eighth international conference on language resources and evaluation, ELRA, pp. 2721–2726 (2012)
21. Schwartz, H. A., Park, G. J., Sap, M., Weingarten, E., Eichstaedt, J. C., Kern, M. L., Ungar, L. H.: Extracting human temporal orientation from facebook language. HLT-NAACL, pp. 409–419 (2015)
22. Wen, M., Zheng, Z., Jang, H., Xiang, G., Rosé, C. P.: Extracting events with informal temporal references in personal histories in online communities. ACL, vol. 2, pp. 836–842 (2013)
23. Kim, S. N., Medelyan, O., Kan, M. Y., Baldwin, T.: Automatic key-phrase extraction from scientific articles. Language resources and evaluation, vol. 47, no. 3, pp. 723–742 (2013)
24. Yih, W. T., Goodman, J., Carvalho, V. R.: Finding advertising key-words on web pages. In: Proceedings of the 15th international conference on World Wide Web, ACM, pp. 213–222 (2006)
25. Mughaz, D.: Classification of hebrew texts according to style. Thesis (in Hebrew), Bar-Ilan University (2003)
26. Koppel, M., Mughaz, D., Akiva, N.: CHAT: A system for stylistic classification of hebrew-aramaic texts. In: The 3th Workshop on Operational Text Classification Systems (OTC-03), vol. 27 (2003)
27. Koppel, M., Mughaz, D., Akiva, N.: New methods for attribution of rabbinic literature. Hebrew Linguistics, A Journal for Hebrew Descriptive, Computational, Applied Linguistics, University Press, vol. 57 (2006)
28. Koppel, M., Mughaz, D., Schler, J.: Text categorization for authorship verification. In: 8th International Symposium on Artificial Intelligence and Mathematics (2004)
29. Liebeskind, C., Dagan, I., Schler, J.: Statistical thesaurus construction for a morphologically rich language. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics, Proceedings of the main conference and the shared task, Proceedings of the Sixth International Workshop on Semantic Evaluation, pp. 59–64 (2012)
30. Liebeskind, C., Dagan, I., Schler, J.: Semi-automatic construction of cross-period thesaurus. LaTeCH'13, vol. 29 (2013)
31. Liebeskind, C., Dagan, I., Schler, J.: Semiautomatic construction of cross-period thesaurus. Journal on Computing and Cultural Heritage (JOCCH), vol. 9, no. 4, pp. 22 (2016)
32. Boyack, K. W., Small, H., Klavans, R.: Improving the accuracy of co-citation clustering using full text. Journal of the American Society for Information Science and Technology (JASIST), vol. 64, no. 9, pp. 1759–1767 (2013)
33. Athar, A., Teufel, S.: Context-enhanced citation sentiment detection. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies Association for Computational Linguistics, ACL, pp. 597–601 (2012)
34. Powley, B., Dale, R.: Evidence-based information extraction for high accuracy citation and author name identification. RIAO'07, pp. 618–632 (2007)
35. Ritchie, A., Teufel, S., Robertson, S.: Using terms from citations for IR: some first results. In: European Conference for Information Retrieval (ECIR), pp. 211–221 (2007)