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[1,3]

[1] Dept. of Computer Science, Bar-Ilan University, 52900 Ramat-Gan, Israel
[2] Dept. of Computer Science, Jerusalem College of Technology, 91160 Jerusalem, Israel
[3] Dep. of Informatics, Kings College London, Strand, London, WC2R 2LS, UK
myghaz@gmail.com, kerner@jct.ac.il, dov.gabbay@kcl.ac.uk

Abstract

In this research, we identify the era in which the author of the given document(s) lived. For rabbinic documents written in Hebrew-Aramaic, which are usually undated and do not contain any bibliographic section, this problem is important. The aim of this research is to find in which years an author was born and died, based on his documents and the documents of dated authors who refer to the author under discussion or are mentioned by him. Such estimates can help determine the time era in which specific documents were written and in some cases identify an anonymous author. We have formulated various types of "iron-clad", heuristic and greedy constraints defining the birth and death years. Experiments applied on corpora containing texts authored by 12 and 24 rabbinic authors show reasonable results.

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

1 Introduction

Citations have great potential to provide important information to researchers in various domains such as academic, legal and religious. Thus, automatic extraction and analysis of citations is growing rapidly and gaining momentum. Computerized corpora and search engines enable accurate extraction of citations. As a result, citation analysis has an increased importance.

Garfield (1965) proposes automatic production of citation indexes, extraction and analysis of citations from corpora of academic papers. Berkowitz and Elkhadiri (2004) extract author names and titles from documents. Giuffrida et al. (2000) use a knowledge-based system to extract metadata including author names from computer science journal papers. Seymore et al. (2006) use hidden Markov models for author name extraction from a narrow corpus of computer science research papers.

Tan et al. (2006) present an approach to author disambiguation for the results of automatically-crafted web searches. 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).

Improvement of retrieval performance using terms has been performed. Bradshaw (2003) uses terms from a fixed window round citations. Dunlop and van Rijsbergen (1993) use the abstracts of citing papers. Ritchie et al. (2007) show that document indexing based on terms combinations used by citing documents and terms from the document itself give better retrieval performance than standard indexing of the document terms alone. Ritchie et al., (2008) investigate how to select text from around the citations in order to extract good index terms in order to improve retrieval effectiveness.

Other researchers have solved various temporal citation-related problems, on the traditional Western scientific literature. Popescul et al. (2000) introduce a method for clustering and identifying temporal tends in hyper-linked scientific document databases.

Citations are defining features not just of academic papers but also of rabbinic responsa (answers written in response to Jewish legal questions authored by rabbinic scholars). Citations included in rabbinic literature are more complex to define and to extract than citations in academic papers written in English because: (1) There is no reference list at the end of a responsa; (2) There is an interaction with the complex morphology of Hebrew and Aramaic; (3) NLP in Hebrew and Aramaic has been relatively little studied; (4) Many citations in Hebrew-Aramaic documents are ambiguous; and (5) At least 30 different syntactic styles are used to present citations (HaCohen-Kerner et al., 2011).

Hebrew-Aramaic documents present various interesting problems: (1) Hebrew is richer in its morphology forms than English. Hebrew has 70,000,000 valid (inflected) forms while English has only 1,000,000 (Berkowitz and Elkhadiri, 2006). Hebrew has up to 7000 declensions for one stem, English has only a few declensions; (2) these types of documents include a high rate of abbreviations

(about 20%) (Dunlop and van Rijsbergen, 1993); and (3) these documents include a high rate of undated citations (HaCohen-Kerner et al., 2011). HaCohen-Kerner et al. (2011) applied six machine learning methods for automatic processing of Hebrew-Aramaic documents, and identification of citations in them. Their research identified whether a sentence includes a citation, though did not identify the citation itself.

There have been many studies on citations in information retrieval (e.g. IR (Berkowitz and Elkhadiri, 2004; Dunlop and van Rijsbergen, 1993; Powley and Dale, 2007; Ritchie et al., 2007; Ritchie et al., 2008; Wintner, 2004; Athar and Teufel, 2012; Kevin et al., 2013)). However, our research is unique in addressing the much more difficult problem of citations included in rabbinic literature. HaCohen-Kerner and Dror Mughaz (2010) present the first citation-based approach to date undated authors. Their experiment was based on a small corpus containing 3,488 documents authored by only 12 authors.

In this research, various extensions are presented: there are two corpora: one containing 10,512 composed by 12 scholars and the second containing 15,450 composed by 24 scholars; there is a use of years that are mentioned in the text documents; constants to the greedy constraints were added; new rabbi's constrains were formulated; and two new manipulations, "Current Year" and "Age" were applied.

This research presents a model that estimates the date of undated documents using undated citations of other dated authors who refer to him or mentioned by him. The estimations based on various constraints of different degree of certainty: "iron-clad", heuristic and greedy. The constraints are based on general citations without cue words and citations with cue words such as: "master", "friend", and "late" ("of blessed memory").

This paper is organized as follows: Section 2 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 3 describes the model. Section 4 introduces the tested dataset, the results of the experiments and their analysis. Section 5 summarizes, concludes and proposes future directions.

2 Citation-Based Constraints

This section presents the citation-based constraints formulated for the estimation of the birth and death years of an author X based on his documents and on other authors' (Y_i) documents who mention X or one of his documents. We assume that the death years and birth years of all authors are known, excluding those of the investigated author. Below are given some notions and constants that are used: X – The author under consideration, Y_i – Other authors, B – Birth year, D – Death year, MIN – Minimal age (currently 30) of a rabbinic author when he starts to write his response, MAX – Maximal life period (currently 100) of a rabbinic author and RABBI_DIS – The age distance between rabbi and his student (currently 20). The estimations of MIN, MAX, RABBI_DIS constants are only heuristic, although they are realistic on the basis of typical responsa authors' lifestyle.

Different types of citations exist: general citations without cue words and citations with cue words, such as: "rabbi", "friend", and "late" ("of blessed memory"). There are two kinds of citations: those referring to living authors and those referring to dead authors. In contrast to academic papers, responsa include 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.

2.1 "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 Y_i, 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 Y_i 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 Y_i 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

$$D(X) >= MAX(Y) \quad (3\ (I))$$

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

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 Y_i (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 Y_i refer to X as "late"), we know that all Y_i died after X, but we do not know when they were born in relation to X's birth. Y_1 was born before X's birth; Y_2 was born after X's birth but before X's death; and Y_3 was born after X's death.

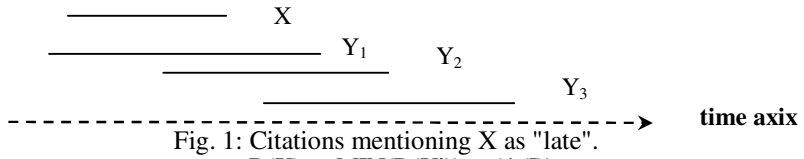

Fig. 1: Citations mentioning X as "late".

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

However, we know that X must have been dead when Y_i 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 Y_i (i = 1, 2, 3) as "late". All Y_i died before X's death (or maybe X is still alive). Y_1 died before X's birth; Y_2 was born before X's birth and died when X was still alive; and Y_3 was born after X's birth and died when X was still alive.

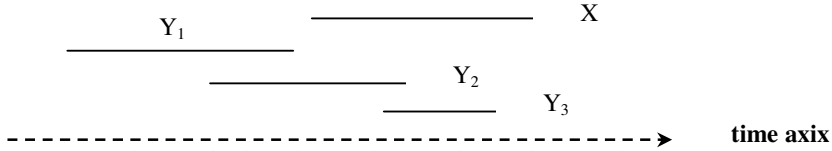

Fig. 2: Citations by X who mentions others as "late".

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

X must be alive after the death of all Y_i 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.

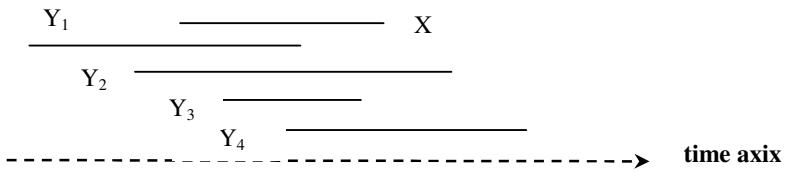

Fig. 3: Citations by authors who refer to X as their Friend/Rabbi.

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

All Y_i 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 Y_i.

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

Again, all Y_i 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 Y_i.

2.2 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 Y_i 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

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

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, minuses MIN is larger than X's born year.

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 Y_i relate to X as dead. Therefore, most of Y_i 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 Y_i and Constraints 15-17 are due to Y_i 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 Y_i died before X's birth; in addition an author doesn't writes 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 no 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 Y_i 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 Y_i 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.

2.3 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 > 2014) { 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, we divide that difference by 2 and normalize D and B to result with an age of 100.

3 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"/"rabbi" 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**.**

4 Experimental Results

The documents of the examined corpus were downloaded from the Bar-Ilan University's Responsa Project[1]. The examined corpora contain 15,450 responsa written by 24 scholars, averaging 643 files for each scholar. These authors lived over a period of 228 years (1786–2014). These files contain citations; each citation pattern can be expanded into many other specific citations by replacing the name of the author and/or his book by one of their other names and abbreviations of their name.

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, based on an extension process using regular expressions.

We divide the data into two sets of authors documents (1) 12 scholars: containing 10,512 files, on average 876 files for each scholar spread over 134 years (1880–2014); (2) 24 scholars: containing

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

15,450 files, on average 643 files for each scholar spread over 228 years (1786–2014) (the set of 24 authors contains the group of 12 authors).

Since this is a novel problem, it is difficult to evaluate the results in the sense that although we can compare how close the system guess is to the actual birth/death years, what we cannot do is assess how-close-is-close, i.e. there is no real notion of what a 'good' result is. Currently, we use the notion Distance, which is defined as the estimated value minus the ground truth value.

Each one of the following tables presents results for each experiment: Greedy (sub-section 2.2) and Iron+Heuristic (sub-section 2.1). Every algorithm was performed on two groups of authors: A group of 12 authors and a group of 24 authors. For every algorithm execution there are results containing estimated dates of birth and death. The results shown in the table are the best birth/death date deviation results (the lower the deviation the better the result). In each quarter of the there are four columns: deviation without refinement, deviation with "Late" refinement, deviation with "Master/Rabbi" refinement, and deviation with "Friend" refinement (beginning of Section 2). In addition, we used two manipulations: Age and Current year (sub-section 2.3). The highlighted bold cells are the cells with the best results. There are four tables: using constants, years, both and neither (beginning of Section 2). In short: There are 4 tables and each table contains 32 results; in all 128 results.

The Age manipulation gives the best results with 94.5% for all refinements, in both algorithms, with or without constants (121/128=0.945; 121 of 128 results were achieved using the Age manipulation and in a few cases adding use of the current year manipulation). It doesn't necessarily mean that for all the authors this manipulation was done, but for some of them it was necessary. The reason this manipulation is effective is because it is only used when there is an age anomaly. Such anomalies occur when the birth year or death year estimate is erratic; therefore, a manipulation is necessary and usually improves the results. For example: In the Greedy with the use of years with the "late" refinement for 24 authors, the estimated birth year and death year are improved for 18 authors each. In short: The Age manipulation is critical for birth and death year reckoning.

	# of authors	No refinemen	Late	Master	Friend	No refinement	Late	Master	Friend
Iron + Heuristic	12	Age	Age	Age	Age	Age	Age	Age	Age
		26.5	50.23	27.5	**12.79**	24.33	**16.77**	24.5	18.21
	24	Age	Age	Age	Age	Age	Age	Age	Age
		34.25	40.73	32.06	**22.67**	24	**15.04**	25.94	21.04
Greedy	12	Age	Age	Age	Age	Age	Age	Age	Age
		16.67	29.21	21.17	17.79	16.08	17.21	20.58	**12.38**
	24	Age	Age	Age	Age	Age	Age	Age	Age
		22.23	27.5	24.31	**20.29**	27.98	**20.83**	29.23	24.33

Table 1: Birth average distance Table 2: Death average distance without constant and without years

	# of authors	No refinemen	Late	Master	Friend	No refinemen	Late	Master	Friend
Iron + Heuristic	12	Age	Age	Age	Age	Age	Age	Age	Age
		50.67	50.27	40.08	**12.75**	**9.67**	16.81	17.33	18.25
	24	Age	Age	Age	Age	no tuning	Age	no tuning	Age
		53.67	40.73	41.13	**24.35**	**13.08**	15.04	23.63	22.73
Greedy	12	Age	Age	Age	Age	Age	Age	Age	Age
		51.5	29.21	41.29	**17.79**	**10.5**	17.38	19.04	12.38
	24	Age	Age	Age	Age	Age	Age	Age	Age
		49.42	27.5	38.79	**23.5**	23.29	**20.83**	28.46	25.92

Table 3: Birth average distance Table 4: Death average distance without constant and with years

Examination of references to years effect indicate that the contribution of years references leads to improve (of the two algorithms with refinements) of 2.76 years in average estimating the Death year. This more prominent Iron+Heuristic (average improvement of 4.16) than with Greedy (average improvement of 1.35). The main reason for that is that an author usually writes until close to his death. Also when a year is mentioned in the text, most of the time is the current year in which he writes his ruling. Since that, the maximum year mentioned in his writings is close to his death year.

Contrast with death year assessment, born year assessing has a negative impact (in the two algorithms with refinements), the deviation increase in average of 10.42. An analysis of the formulas shows that the formula determines the birth year in the Greedy (10(G)) uses the most recent year the author mentioned. The most recent year the arbiter mentioned is usually near his death as explained above, thus very poor birth outcomes, deterioration of 12.48 years. The results of the Greedy are better than Iron+Heuristic (deterioration of 8.36 years), but the effect of references of years on the results of Iron+Heuristic is less harmful (as explained earlier). In conclusion, in order to estimate the death year we will execute the Iron+Heuristic algorithm, we will use references to years without any refinement.

	# 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	47.81	27.17	**12.79**	22.17	**18.02**	23.83	18.21
	24	Age	Age	Age	Age	Age	const	Age	Age
		30.42	40.93	33.35	**22.17**	20.67	**18.19**	21.21	22.75
Greedy	12	const	Age	const	Age	Age	Age	Age	Age
		16.83	22.29	13.67	**13.42**	44.08	24.46	43.04	**17.08**
	24	Age	Age	Age	Age	Age	Age	Age	Age
		14.31	18.25	15.94	**14.13**	42.15	26.04	40.19	**22.71**

Table 5: Birth average distance Table 6: Death average distance with constant and without years

The “friend” refinement for birth year estimation gives the best results in comparison with the other refinement options - “late”, “master” or none. This is due to friends being of the same generation and more or less the same age, thus are born roughly the same year. Therefore, an author referring to another author as his friend - the estimate of his birth year will ensue good results. For the death year, however, this is not assured since there may be a much greater period between the deaths of friends. (One author may pass away at the age 50, while his friend at the age 75). Hence the "friend" refinement usually ensues better results for birth year assessment than for death year assessment.

After we discovered that the best results for the birth year always with “friend” refinement (except for one case in the Greedy without years, constants or any refinement, for 12 authors), we examined at greater depth and found that this occurs specifically with use of constants. Use of constants is important: it results with an average improvement of 6.29 years in the Greedy (for 12 and 24 authors). In general, reference to a Posek in Responsa is only after he becomes important enough to be mentioned and regarded in Halachic Responsa that usually at advanced age.

We mentioned above that the Greedy with the use of constants presents the greatest improvement. Even without the use of constants the Greedy gives the best results. The reason lies in the formula, the formula (13(G)) finds the lowest birth year from the group of authors that the arbiter mentioned (we chose a constant of 20). Unlike the formula of the Greedy, the formula of Iron+Heuristic (7(H)) reduces constant and therefore the results of the Greedy are better. In conclusion, in order to best assess the birth year we apply the Greedy algorithm, use with constant and "friend" refinement.

	# of authors	No refinement	Late	Master	Friend	No refinement	Late	Master	Friend
Iron + Heuristic	12	Age	Age	Age	Age	Age	const	Age	Age
		31.42	30.58	31.17	**16.58**	**9.17**	11.92	16.92	17.67
	24	Age	Age	Age	Age	const	const	Age	Age
		35.06	40.93	35.69	**22.6**	**13.08**	18.23	20.04	24.44
Greedy	12	Age	Age	Age	Age	Age	Age	Age	Age
		21.5	21.21	18.63	**13.42**	29.08	29.38	35.96	**17.08**
	24	Age	Age	Age	Age	Age	Age	Age	Age
		21.29	18.25	20.35	**17.33**	32.88	26.04	37.56	**24.29**

Table 7: Birth average distance Table 8: Death average distance with constant and with years

Now we examine what the usefulness of the combination of using constants with references to years. The best results evaluating birth year is with the Greedy using constants and without using references to years. The best results evaluating death year is with the Iron+Heuristic using constants and without references to years. When we compare these results with the results shown in tables 7 and

8 we find that: In the Greedy there is an improvement in only one case, 12 authors using "late" finesse and in the rest there decrease or stability; In the Iron+Heuristic there decrease in two results and improvement in 5 results. In the Iron+Heuristic there an average improvement of 0.64 years, and in fact, the best result death year estimation. Possible cause that reduces the quality of the results of the Greedy is that estimation of birth year using references to years impairs pretty severe the results (explained above); because the effect of references to years the improvement that constants does, cannot cause the combination of them enhance the results. In contrast, assessing death year, using references to years with Iron+Heuristic significantly improves the results and using constants improves it a little more; therefore, combination of constants + years brings better results assessing death year. In conclusion: To assess death year we use references to years and constants by the Iron+Heuristic algorithm; To evaluate birth year we'll run the Greedy algorithm with the use of constants and "friend" refinement without the references to years.

In this section we present the differences between the two algorithms from the perspective of the results. Each table presents the results of the two algorithms when we check which of them gives the better results we get that the Iron+Heuristic gives the best results in the estimation of death year. When using references to years, concerning death year, the results assessment of Iron+Heuristic vastly improved. When we are not using a reference of years and without the use of constants, the average results of death year estimation, with Greedy and the Iron+Heuristic are quite close (with Iron+Heuristic 21.23 , with Greedy 21.08). When using constants, results of the Iron+Heuristic assessing death year better than the Greedy. We can see from the table that using constants, in the Iron+Heuristic, does not change the outcome assessment of death year significantly (average 20.63); although, the combination of using constants and using references to years improves the results (average 16.43). Explanation of anything said in this paragraph appears in the preceding paragraphs. In conclusion: in order to assess the death year we should use the Iron+Heuristic using reference of years and without the use of any refinement (average 17.07), the addition of Using constants improves slightly the results (average 16.43).

In the Greedy there is a different phenomenon, the effect of using the constants impairs the quality of the results, and the average results are decrease by over 50% (from average of 21.08, for all eight results to of 32.47). For example: the Estimation of the death year of the late Rabbi Ovadia Yosef has an error of 60 years (instead of 2013 the algorithm result is 1953), determining he died at age of 33. According to the formula, when using constants we reduce constant (currently 30), then the error increased and therefore, of the Greedy results assessing death year decreases. For example death year assessment of Rabbi Ovadia Yosef, instead the true death year which is 2013, the algorithm result is 1953, 60 years error, with the use of constant the error is 90. The reasons which written here shows the Greedy does not evaluate death year good and constants decrease them. In contrast, the assessments the Greedy gives the best results for birth year (as explained above). General Conclusions: In order to evaluate birth year we will use Greedy algorithm using constants and "friend" refinement. To assess death year we will use Iron+Heuristic algorithm with mentions of years and without any refinement.

5 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 formulate various kinds of iron-clad, heuristic and greedy constraints.

The best estimates of birth year have been achieved using the Greedy version with the use of constants using "friend" refinement. The best assessment of death year has been achieved using the Iron+Heuristic version with the use of constant and year without any refinement.

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

Awais Athar, and Simone Teufel. "Context-enhanced citation sentiment detection." Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2012. 597-601.

Eric Berkowitz and Muhammed Reda Elkhadiri. "Creation of a style independent intelligent autonomous citation indexer to support academic research." 2004. 68-73.

Kevin W. Boyack, Henry Small, and Richard Klavans. "Improving the accuracy of co-citation clustering using full text." Journal of the American Society for Information Science and Technology 64(9) (2013): 1759-1767.

Shannon Bradshaw. "Reference directed indexing: Redeeming relevance for subject search in citation indexes." *Research and Advanced Technology for Digital Libraries*. Springer Berlin Heidelberg, 2003. 499-510.

Mark D. Dunlop *and Cornelis J.* Van Rijsbergen. 1993. "Hypermedia and free text retrieval."*Information processing & management* 29(3): 287-298.

Eugene Garfield. 1965. "Can citation indexing be automated." *Statistical association methods for mechanized documentation, symposium proceedings*. 189-192

Giovanni Giuffrida, Eddie C. Shek, and Jihoon Yang. 2000. "Knowledge-based metadata extraction from PostScript files. " Proceedings of the fifth ACM conference on Digital libraries. ACM. 77-84.

Yaakov HaCohen-Kerner, Ariel Kass, and Ariel Peretz. 2008. "Combined one sense disambiguation of abbreviations." *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics. 61-64.

Yaakov HaCohen-Kerner and Dror Mughaz. "Estimating the birth and death years of authors of undated documents using undated citations." Advances in Natural Language Processing. Springer Berlin Heidelberg, 2010. 138-149.

Yaakov HaCohen-Kerner, Nadav Schweitzer, and Dror Mughaz. 2011. "Automatically Identifying Citations in Hebrew-Aramaic Documents." *Cybernetics and Systems: An International Journal* 42(3): 180-197.

Alexandrin Popescul, Gary W. Flake, Schulman Lawrence, Lyle H. Ungar and C. Lee Giles. 2000. "Clustering and identifying temporal trends in document databases." Advances in Digital Libraries, 2000. Proceedings. IEEE: 173-182.

Brett Powley, and Robert Dale. 2007. "Evidence-based information extraction for high accuracy citation and author name identification." *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE. 618-632.

Anna Ritchie, Simone Teufel, and Stephen Robertson. 2008. "Using terms from citations for IR: some first results." *Advances in Information Retrieval*. Springer Berlin Heidelberg: 211-221.

Anna Ritchie, Stephen Robertson, and Simone Teufel. 2008. "Comparing citation contexts for information retrieval." *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM. 213-222.

Kristie Seymore, Andrew McCallum, and Roni Rosenfeld. 1999. "Learning hidden Markov model structure for information extraction." *AAAI-99 Workshop on Machine Learning for Information Extraction*. 37-42.

Yee F. Tan, Min Y. Kan, and Dongwon Lee. 2006. "Search engine driven author disambiguation." *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. ACM. 314-315.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. "Automatic classification of citation function." *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 103-110.

Shuly Wintner. 2004. "Hebrew computational linguistics: Past and future." *Artificial Intelligence Review* 21(2): 113-138.