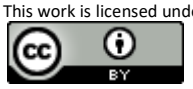
Available online at: https://ijact.in

Date of Submission	18/01/2019
Date of Acceptance	15/02/2019
Date of Publication	28/02/2019
Page numbers	3031-3035 (5 Pages)



ISSN:2320-0790

An International Journal of Advanced Computer Technology

REASONING ABOUT INEXACT DATES USING DENSE VECTOR REPRESENTATION

Davor Lauc
University of Zagreb
davor@lauc.org

Abstract: Representation and reasoning with temporal data is a well-researched problem in logic and computer science. Although many practical applications need the representation of inexact dates and reasoning with such representations, there is no standard developed methodology for it. In this paper, we propose a standard representation of inexact dates based on discrete probability distributions. Inspired by recent breakthroughs in natural language processing and information retrieval in embedding words as dense vectors we have developed a similar approach for representation and comparison of inexact dates.

Keywords: Temporal logical reasoning, Temporal data representation, Inexact data, Word embedding, Information retrieval.

I. INTRODUCTION

The problems related to the representation of temporal information and reasoning with them has puzzled researcher in many fields of sciences, engineering and humanities for ages. From ancient philosophers' inquiries on the nature of time [1][2] to contemporary research of temporal reasoning within digital humanities [3][4] and artificial intelligence [5]a large number of theories, methodologies, and formalism are developed to deal with various aspects of time. The whole academic subfields, like chronology and periodisation within history, or tense logic within philosophy, have been devoted to this problem.

In this paper, we focus on one aspect of temporal reasoning, the digital representation of inexact dates that enables efficient reasoning in the context of information retrieval and record linkage. Although the developed system is applicable in other all where inexact temporal information is present, the focus of this research is on developing the effective method to embed uncertain dating of events related to persons in genealogical records and to infer relationships among them.

Information on the past dates in historical records are often imprecise, and frequently only a vague range of temporal values can be only guessed. For example, in the baptismal parish registers, one of the typically biggest source for family history, the date of the most important temporal information, the date of the birth, is not explicitly stated for any person named in the record. We can only infer that the child has been born in a short period before the day of the baptism, although it varies from a few days to a year depending on the congregation and historical period[6][7].

We can further guess that it is most probable that mother was at her fertility peak, though the probability distribution of fertility fluctuates in history and geography [8], and so on. Nevertheless, such uncertain inference of temporal data is often vital when researcher conclude from such evidence, like identify all the records that refer to the same person. To be able to automate such inference, the first step is a proper representation of the uncertain temporal data.

II. INEXACT DATES REPRESENTATION

The proper representation of temporal information is a well-researched problem in logic, computer science and artificial intelligence. At the beginning of the second half of the twentieth century, logician Arthur Prior developed temporal logic as an extension of classical logic, with the modal operators such as “It has at some time been the case that…” [9]. The approach inspired by Donald Davidson extension of first-order logic with a temporal argument has inspired James Allen interval algebra [10][11]. To the present days, the number of alternative approaches has been developed, within the fields of temporal databases [12], linked data [13], logic programming [14] and many others.

The common characteristic of all these approaches is that that they represent temporal data as either point or interval. However, in the many contexts, especially within humanities in general and history in particular, temporal intervals have often vague boundaries, and some periods are more probablethan others. For example, if some event took place during the industrial revolution, it is possible that it happened in 1745, but this is less probable than, e.g. 1801. If there is a record of a mother of the baptized child in 1848, it is more probable that she is born in 1828 than in 1814.

Even when some source quotes the exact date, we can put more or less reliance on it, depending on the credence of the source. So, in the case of the information from unreliable sources, we want their representation to be close or more similar to the representation of the surrounding times.

One of the obvious ways to design such representation is to represent dates as a discrete probability distribution over the timeline. In this research the granularity of days is sufficient, but the same representation can be used for more fine-grained granularity such as hours, minutes or seconds.Consequently, the vague temporal period can be represented as a vector of real numbers from the interval [0, 1] that signify the probability that the corresponding day is the real date of the event we are embedding. The size of the vector is the number of days between the first date we need to represent to the present. To satisfy the requirement of discrete probability distribution the sum of all probability masses must add to one.

The reliability of information, or precision, can be expressed as a smaller or higher variance around the mean of the date. In such a way, we can easily describe any temporal determinant, from exact dates, months, years, centuries, to the more complicated qualified times or date ranges.

Figure1 depicts some examples of such representations:

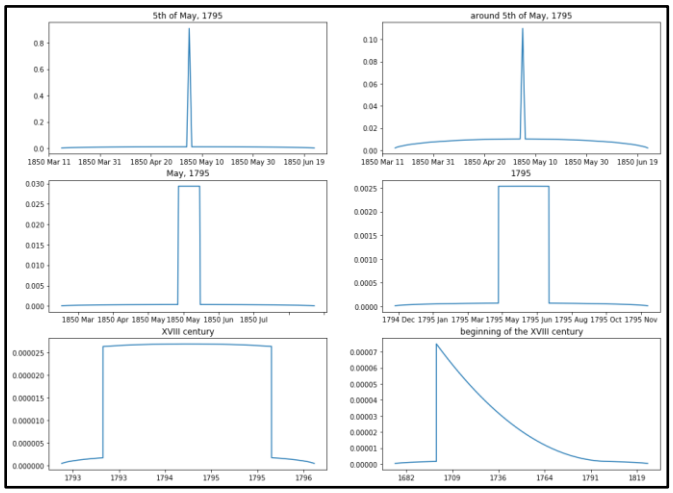


Figure 1

The first graph represents the exact date from a reliable source, allowing for the small spread of the of the 10% probability mass to the surrounding times. The possibility of error is represented by the beta distribution with the parameters $\alpha = 1.5, \beta = 1.5$. The second probability distribution represents the same date, but from a very unreliable source. Although the mean and the highest mass is at the exact date, the 90% of the probability is spread around the date using the same beta distribution.

The third graph represents an example of less precise information, namely any day in May 1795. The month is expressed using the uniform distribution over all days in the month, again allowing some uncertainty by distributing some weight to the close dates. The same principle applies to the fourth and fifth graphs, where former represents a year and the later the whole century.

In some historical sources, a qualifier is applied to the range of dates, like the middle, beginning or the end of some period. Such dateis easily represented using more skewed distribution, like in the last graph depicting the beginning of the 18^{th} century, again using beta distribution but with parameters$\alpha = 1, \beta = 3$.

This representation also facilitates representation of incomplete information thatmethodologies using intervals cannot even approximate. For example, if we know that today is the birthday of some college student, but we do not know her or his age, we still can represent her or his day of birth as an uncommonbut proper probability distribution that assignsprobability mass on today’s date in the years between, say, 17 and 25 years ago.

Frequently, there is no explicit dating of some event, but it is possible to infer temporal determinants from the contextual information in the source. For example, if a source quotes that some person was, for example, a member of Christopher Columbus' crew on the first voyage, in the absence of other information; we can only

deduce that the person was probably born before 1778 and after 1730. In the context of genealogical research, in the baptism records, the dates of births of parents are ordinarily not inscribed. In such cases, it is possible to use available demographic data to find and fit an appropriate probability distribution, and make an educated guess about the implicit dating. For example, if we analyze a baptism record from the 1940s, we can use the following demographic data:

		Age of mother							
Year	**Total**	**Under 15**	**15–19**	**20–24**	**25–29**	**30–34**	**35–39**	**40–44**	**45–49**[2]
1940	2,558,647	3,865	332,667	799,537	693,268	431,468	222,015	68,269	7,558

The empirical distribution is generalized by finding and fitting parameters to some probability distribution. Figure 2 represents the date of birth of a person that has become a mother in 1940 using the beta distribution ($\alpha = 9.5, \beta = 7$), inferred from the empirical statistical data shown by the blue curve.

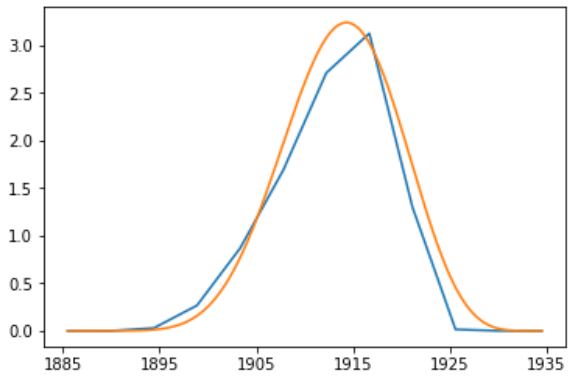


Figure 2

Using this approach, we can express our best guess about any inexact dates. Using standard techniques of Bayesian statistics [15], we can express our uncertainty about temporal information as priors, and learn better representations when empirical distributions are available either from public sources or our data.
As we can represent all inexact temporal information, at least in principle, in this way, the next natural step is to develop a reasoning system for such representations.

III. Reasoning with Inexact Dates Representation

Standard tasks in temporal reasoning include inferring relationship among temporal points and intervals, reasoning about actions and changes, assessing the consistency of a set of temporal information [16][17]. In this paper, we focus on inferring the fundamental relation between two dates – the relation of equivalence. As the dates are inexact, instead of classical binary relation, we need to infer the fuzzy relationship between two dates. In the case of imprecisedates, it is convenient to express the relationship between two periods as a similarity metric. Such similarity metrics can be used as a component in an information retrieval system or a more complicated reasoning system.

IV. Defining Measure of Inexact Dates Similarities

The standard requirements for any metric function are non-negativity, symmetricity, the identity of indiscernibles, and the triangle inequality. Additionally, it is convenient when the metric is normalized so that it takes a real value in the [0,1] interval. As the dates are represented as standard probability distributions, it wouldseem convenient to use some of the existing distance or similarity metrics. There are many such similarity measures and distance function between probability distribution in different scientific fields. In statistics and probability theory there are distance correlations, Bhattacharyya distance f-divergences like Kullback–Leibler, Kolmogorov–Smirnov and so on [18]. The most commonly used measure is Bhattacharyya similarity. For inexact dates represented as random variables X and Y, it would be calculated as:

$$S_{bhatt}\,(X,Y) = \sum_{i=0}^{today} \sqrt{X_i \times Y_i}$$

Another possible applicable distance function may be information theoretical measures like mutual information or Jensen–Shannon divergence, as a symmetric version of Kullback–Leibler divergence [19].
The principal problem with all those distances is that they do not satisfy the basic intuitive semantics of the inexact date'scomparison. Namely, the reason for representing dates as the probability distribution is to express our ignorance about the exact date of some event. The event happened on some exact date; wedo not have enough information of confidence to determine it. As most of the above distance and similarity measures are proper metrics, they satisfy the identity of indiscernibles, so the distance between two identical probability distribution is zero and, consequently, the similarity is maximal (1). This property is not desired in our context as our chosen similarity is the probability that two events have happened on the same day giventhe two inexact dates representation. If we signify with x_e date of the first event and with y_e date of the second one, and with X and Y probability distributions representing our uncertainty about the exact dates when the events occurred, we can define the similarity as a conditional probability:

$$p(x_e = y_e|X, Y)$$

The consequence of such definition is that our desired similarity measure is not proper metrics, as it violates condition that $d(x, y) = 0 \leftrightarrow x = y$. One simple counterexample, where r is inexact date representation using uniform distribution, is

$$\begin{aligned} p(x_e = y_e|r(XIX\ century), r(XIX\ century)) \\ < p(x_e = y_e|r(1848), r(1848)) \\ < p(x_e = y_e|r(1.5.1848), r(1.5.1848)) \end{aligned}$$

The only case where we want that similarity of two representation is one is when there is no uncertainty, and whole probability mass is assigned to a single day. In all other cases, similarity should be significantly lower.

Although it is very hard to calculate exact similarity, because the joint probability distribution of two inexact dates representation is more often than not unknown, it is plausible to use a strong (naive) independence assumptions between probability distribution. With this assumption, the similarity measure is a simple scalar product between two vectors. As those vectors are normalized (they are probability distributions), this measure is equivalent to the cosine similarity, and one of the most commonly used measures in information retrieval.

As we are applying dimensionality reduction to the inexact date representation, and such reduction inevitably adds some noise to the similarity calculation. For example, the probability that two events occurred on the same day if we know that they happened in the same century is only2.737e-05. As we are primarily interestedin preserving relative ranking similarity and not calculating the exact probability, we are using a scaled version of the scalar product, to minimize the error introduced by dimensionality reduction. So the similarity function we are using is:

$$s(X, Y) = \sqrt[4]{X \cdot Y}$$

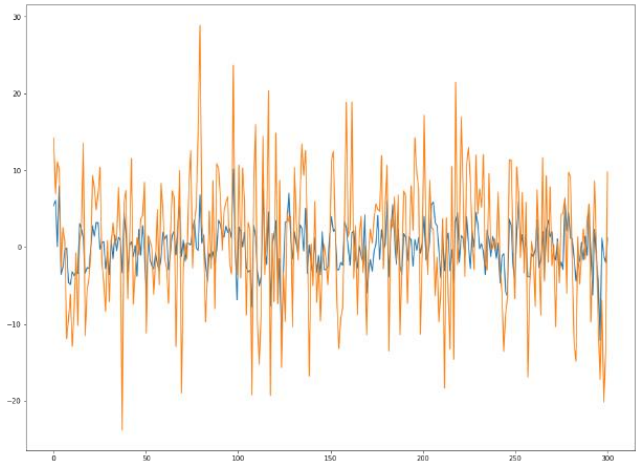
V. DIMENSIONALITY REDUCTION

Dates representations as probability distributions with the day granularity can be very large vectors. When the first possible date,happened in the distant past, it can have a dimensionality of several hundred thousands real numbers. In the case of big datasets, where every entity contains several temporal information, the developed representation can be computationally expensive. As in the almost all real-world usage cases, the spread of possible dates is limited, the most of the dimensions are zero.

This problem is akin to the problems caused by the standard one hot representation of words in the natural language processing system. The recent developments of latent semantical analysis and the dense vector representation of words meaning have provided ground-breaking forays in areas of natural language processing, information retrieval and connected fields [19, 20]. As our focus is on inferring dates similarities, the similar techniques can be used to reduce the dimensionality of dates vectors. An additional benefit of such dense representation is that it is easier to combine dates vectors with word vectors in more complex retrieval and record linkage systems.

To perform dimensionality reduction, Siamese neural network [21]is designed. The input of the network shared is two date representation that is reduced to the dimensionality of 300 with the embedding layer with shared weights. The reduced vectors are fed to the cosine similarity layer that produces output as a real number in [0,1] interval. The loss function is a standard quadratic loss. The network is trained on 100M generated positive and negative dates pairs, where 50% is generated such that the above similarity function has a value greater than 0, and the other half is zero.

With this simple network architecture, satisfactory dimensionality reduction is obtained, with the total quadratic loss less than 0.006. Figure 3 represents an example of reduced vectors the two similar dates (the day 5/1/1750 and the year 1750). Although the semantic transparency is lost, and the distributions are not readable from the dense representation, relation of similarity is preserved.



VI. CONCLUSION AND FURTHER RESEARCH

The developed representation for an inexact date, similarity measure and dimensionality reduction system has been implemented into retrieval and entity resolution system for genealogical data. As the final results are intervened with the other non-temporal date, it is hard to evaluate the contribution of this system. So the natural next step of this research is to develop an evaluation dataset of inexact dates that includes ranked similarity among them, as such dataset, to the author's best knowledge does not exist. Such dataset would facilitate evaluation of this approach in the information retrieval. Construction of such dataset should include the creation of neural models for automatic translation of inexact temporal expression to the structured representation. Some variants of such structured presentation are already proposed in the Wikidata date

format proposal [23] and other semantic web community research projects [24].

The other lines of future research include a more theoretical solid definition of the similarity function, perhaps in the context of belief functions and evidence theory [25]. Further research into the development of the dimensionality reduction neural model architecture and the development of specialized word vectors for inexact dates also seems conceivable.

VII. REFERENCES

[1] Aristotle, Aristotle's Physics, U of Nebraska Press, 1961.

[2] R. L. Poidevin, The Images of Time: An Essay on Temporal Representation, Oxford: Oxford University Press, 2007.

[3] O. &. M. M. F. Kolomiyets, "KUL: Recognition and Normalization of Temporal Expressions," in Proceedings of the 5th International Workshop on Semantic Evaluation, 2010.

[4] A. Rabinowitz, "It's about time: historical periodization and Linked Ancient World Data," ISAW Papers 7.22, 2014.

[5] Y. &. G. N. Shoham, "Temporal reasoning in artificial intelligence," in Exploring artificial intelligence, 1988, pp. 419-438.

[6] B. M. &. S. R. S. Berry, "Age at baptism in pre-industrial England," A Journal of Demography ,pp. Volume 25, 1971 - Issue 3, 1971.

[7] J. &. D. R. Boulton, "Few deaths before baptism: clerical policy, private baptism and the registration of births in Georgian Westminster: a paradox resolved," Local population studies 94.1 ,pp. 28-47., 2015.

[8] S. e. Greenhalgh, Situating fertility: Anthropology and demographic inquiry., Cambridge University Press, 1995..

[9] A. N. Prior, Time and Modality, Oxford: Clarendon Press, 1957.

[10] D. Davidson, "The Logical Form of Action Sentences," in The Logic of Decision and Action, Pittsburgh, University of Pittsburgh Press, 1967, p. 81–95.

[11] J. F. Allen, "Maintaining knowledge about temporal intervals," Communications of the ACM. ACM Press, p. 832–84, 1983.

[12] N. T. B. K. I. &. T. Y. Pelekis, "Literature review of spatio-temporal database models.," The Knowledge Engineering Review, 19(3), pp. 235-274, 2004.

[13] G. a. S. M. a. M. I. a. S. N. Correndo, "Linked timelines: temporal representation and management in linked data," in Proceedings of the First International Conference on Consuming Linked Data - Volume 665, Aachen, Germany, 2010.

[14] C. a. a. Brenton, "Answer Set Programming for Qualitative Spatio-Temporal Reasoning: Methods and Experiments.," OASIcs-OpenAccess Series in Informatics. Vol. 52. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[15] A. O'Hagan and M. West, The Oxford Handbook of Applied Bayesian Analysis, Oxford: OUP, 2010.

[16] M. G. D. &. L. V. Fisher, Handbook of Temporal Reasoning in Artificial Intelligence, Elsevier B.V, 2005.

[17] O. (. Stock, Spatial and Temporal Reasoning, KLUWER ACADEMIC PUBLISHERS, 1997.

[18] S.-H. Cha, ""Comprehensive survey on distance/similarity measures between probability density functions," City 1.2, p. 1, 2007.

[19] T. e. a. Mikolov, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems. , 2013..

[20] B. a. N. C. Mitra, "Neural Text Embeddings for Information Retrieval.," in Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. , 2017.

[21] S. H. R. &. L. Y. Chopra, "Learning a similarity metric discriminatively, with application to face verification.," Computer Vision and Pattern Recognition 1, p. 539–546., 2005..