Determining Relevance of Imprecise Temporal Intervals for Cultural Heritage Information Retrieval

Tomi Kauppinen[*,a], Glauco Mantegari[b], Panu Paakkarinen[a], Heini Kuittinen[a], Eero Hyvönen[a], Stefania Bandini[c]

[a]*Semantic Computing Research Group (SeCo), The Aalto University School of Science and Technology and University of Helsinki, Finland*
[b]*QUA_SI Doctoral and Advanced Research Program on the Information Society, University of Milano-Bicocca, Italy*
[c]*Complex Systems and Artificial Intelligence Research Centre (CSAI),University of Milano-Bicocca, Italy*

Abstract

Time is an essential concept in cultural heritage applications. Instances of temporal concepts such as time intervals are used for the annotation of cultural objects and also for querying datasets containing information about these objects. Hence it is important to match query and annotation intervals by examining their similarity or closeness. One of the problems is that in many cases time intervals are imprecise. For example, the boundaries of the "Pre-Roman age" and the "Roman age" are inherently imprecise and it may be difficult to distinguish them with clear-cut intervals. In this paper we apply the fuzzy set theory to model imprecise time intervals in order to determine relevance of the relationship between two time intervals. We present a method for matching query and annotation intervals based on their weighted mutual overlapping and closeness. We present 1) methods for calculating these weights to produce a combined measure and 2) results of comparing the combined measure with human evaluators as a case study. The case

---

[*]Corresponding author. Address: Department of Media Technology, PL 5500, The Aalto University School of Science and Technology, 02150 TKK, Finland. Tel:+35804513355. Fax: +35804513356. Mobile: +358407156945.

Email addresses: `tomi.j.kauppinen@gmail.com` (Tomi Kauppinen), `glauco.mantegari@disco.unimib.it` (Glauco Mantegari), `ppaakkar@cc.hut.fi` (Panu Paakkarinen), `hmkuitti@cc.hut.fi` (Heini Kuittinen), `eero.hyvonen@tkk.fi` (Eero Hyvönen), `stefania.bandini@csai.disco.unimib.it` (Stefania Bandini)

study takes into consideration archaeological temporal information, which is in most cases inherently fuzzy, and therefore offers a particularly complex and challenging scenario. The results show that our new combined measure that utilizes different weighted measures together in rankings, performs the best in terms of precision and recall. It should be used when ranking annotation intervals according to a given query interval in cultural heritage information retrieval. Our approach intends to be generalizable: overlapping and closeness may be calculated between any two fuzzy temporal intervals. The presented procedure of using user evaluation results as a basis for assigning weights for overlapping and closeness could potentially be used to reveal weights in other domains and purposes as well.

Key words: fuzzy sets, time, information retrieval, cultural heritage

1. Introduction

Time is one of the central concepts in ontologies representing the world, and hence should also be centric in annotations and queries of the Semantic Web. Time is especially important for managing historical collections, for example in visualizing them on a timeline (Hyvönen et al., 2009; Schreiber et al., 2006).

However, representing time in Semantic Web ontologies is not straightforward because the question of when a certain time was or will be is often uncertain, subjective or vague (Nagypál and Motik, 2003). For example, it may not be known when exactly a given archaeological artifact was manufactured (uncertainty), when "The Middle Ages" was according to opinions of different historians (subjectivity), or when the spring starts (vagueness, imprecision).

In addition, transitions between different phases, such as historical periods, are usually complex processes which are not identifiable by clear cut dates, even if conventional calendric markers are mostly used in order to simplify historical sequences. All these elements are at the basis of imprecise temporal representations especially in the cultural heritage, historical and archaeological contexts.

Nevertheless, representations of time are needed for representing and matching annotations and queries. The definition of temporal intervals and their relations are crucial in the context of archaeological chronologies. Checking whether two time intervals have something in common allows for answer-

ing queries like "find all artifacts manufactured around the middle of the 1st century B.C.". At the same time, the often inherently fuzzy nature of historical and archaeological temporal information makes this scenario particularly complex and challenging.

In this paper we adapt the idea (Nagypál and Motik, 2003; Visser, 2004) of using fuzzy sets (Zadeh, 1965) for the representation of temporal annotations and queries. The structure of the paper is as follows. In Section 2 we present a new method for matching queries and annotations using their weighted mutual overlapping and closeness. In Section 3 we present a case study where the method has been implemented and provide results of its evaluation with human test subjects. The results in Section 5 suggest that a measure that combines different single, weighted measures together performs best in terms of precision and recall, and should hence be used when ranking annotation intervals with respect to a given query interval. In Section 6 we provide a discussion about the results, and Section 7 provides discussion about the related work. Finally, Section 8 concludes the paper.

2. Representing Fuzzy Temporal Intervals

2.1. Representing and Reasoning about Temporal Overlappings

To represent and calculate overlappings between temporal intervals we use fuzzy sets (Zadeh, 1965) to represent temporal intervals, resulting in fuzzy temporal intervals (Nagypál and Motik, 2003; Visser, 2004). The fuzzy set theory enables modeling imprecise time ranges, such as "around 1950" that have vague boundaries. In the fuzzy set theory the grade of membership μ of the item x in the given set A is a value in range [0,1], whereas in the traditional set theory an item x either belongs to a given set A or not. In other words, in the fuzzy set theory x more or less belongs to the set A.

Following Visser (2004) we define a fuzzy temporal interval T that represents imprecision of a time period. The fuzzy temporal interval T is a quadruple $\langle T_{fuzzybegin}, T_{begin}, T_{end}, T_{fuzzyend} \rangle$, where $T_{fuzzybegin}$ is used to explicate the earliest start of that period, T_{begin} for latest start, T_{end} for the earliest end and finally $T_{fuzzyend}$ for the time when the time period has ended for sure. Figure 1 shows an example of a fuzzy temporal interval of Equation (1) modeled using fuzzy sets. It is a subjective expression of the period "from around the beginning of the I century B.C. to the first half of the I century A.D.". As one can observe, there are 10 years from $T_{fuzzybegin}$ to T_{begin} and 14 years from T_{end} to $T_{fuzzyend}$.

$$
\begin{aligned}
T_{fuzzybegin} \;\; &= 105 \;\; B.C. \\
T_{begin} \;\; &= 95 \;\; B.C. \\
T_{end} \;\; &= 43 \;\; A.D. \\
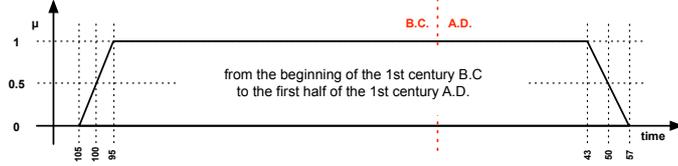T_{fuzzyend} \;\; &= 57 \;\; A.D.
\end{aligned}
\tag{1}
$$

Figure 1: The period "from around the beginning of the I century B.C. to the first half of the I century A.D." represented as a fuzzy temporal interval.

We will also make use of Left-Right notation (LR) (Dubois and Prade, 1988) because of its' suitability for arithmetic operations. The example in Figure 1 in LR-notation is given in Equation (2).

$$
\begin{aligned}
T_{LR} \;\; &= (T_{begin}, T_{end}, T_{begin} - T_{fuzzybegin}, T_{fuzzyend} - T_{end})_{LR} \\
&= (95 \;\; B.C., 43 \;\; A.D., 10, 14)_{LR}
\end{aligned}
\tag{2}
$$

We use the intersection of two fuzzy intervals in order to calculate how much these two intervals proportionally overlap. Proportional overlap function

$$
o_t : A, Q \to p \in [0, 1] \tag{3}
$$

tells how much two fuzzy intervals A and Q overlap. It is represented in terms of temporal overlap $o_t = overlaps(A, Q) = |A \cap Q|/|A|$. Hence the overlap function answers the question "How much a given query interval Q *overlaps* with an annotation interval A?".

Similarly overlappedBy function answers the question "How much a given query interval Q is *overlapped by* an annotation interval A?". Hence we define the proportional overlappedBy function as $o_b : A, Q \to p \in [0, 1]$ and represent it as $o_b = overlappedBy(A, Q) = |A \cap Q|/|Q|$.

4

The universe X in our case is defined to be infinite, which means that $|A|$ i.e. the cardinality[1] of A is defined by $|A| = \int_x \mu_A(x)dx$. Hence we get:

$$o_t = overlaps(A, Q) = |A \cap Q|/|A| = \frac{\int_x \mu_{A \cap Q}(x)dx}{\int_x \mu_A(x)dx} \tag{4}$$

Calculating o_t (or o_b) intuitively amounts to computing and dividing the integral areas of the two membership functions in the formula.

Figure 2 depicts a case where the fuzzy temporal interval Q="Roman age" intersects with another fuzzy temporal interval A="Pre-Roman age". These were modeled subjectively[2] by a domain expert as follows:

$$\begin{aligned} A_{fuzzybegin} &= 510 \ B.C. \\ A_{begin} &= 490 \ B.C. \\ A_{end} &= 222 \ B.C. \\ A_{fuzzyend} &= 89 \ B.C. \end{aligned} \tag{5}$$

$$\begin{aligned} Q_{fuzzybegin} &= 222 \ B.C. \\ Q_{begin} &= 89 \ B.C. \\ Q_{end} &= 452 \ A.D. \\ Q_{fuzzyend} &= 569 \ A.D. \end{aligned} \tag{6}$$
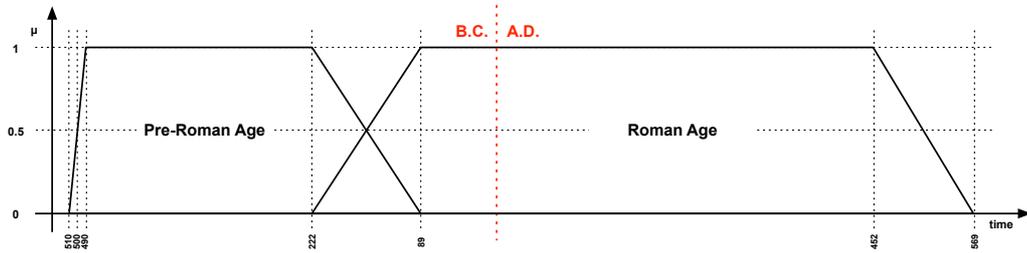

Figure 2: A fuzzy temporal interval "Pre-Roman age" intersects with another fuzzy temporal interval "Roman age".

---

[1]See Zimmermann (1996), page 16, for a discussion about cardinalities of fuzzy sets.

[2]The criteria for the representation of the fuzzy dates represented in the figure are a personal choice of a domain expert, and are related to the specific chronology of the ancient history of the case study city.

As an example, the results of calculation of the values of the functions *overlaps* and *overlappedBy* between time intervals Q="Roman age" and A="Pre-Roman age" are given below.

The value for *overlaps* is obtained by calculating

$$o_t = overlaps = \frac{|A \cap Q|}{|A|} = 0.0974\ldots \approx 0.1 \tag{7}$$

Intuitively speaking, this means that "Roman age" overlaps around 10 percent of "Pre-Roman age". Similarly, the same intersection $|A \cap Q|$ confirms to around 5% of the other period "Pre-Roman age" (denoted with A) and hence

$$o_b = overlappedBy = \frac{|A \cap Q|}{|Q|} = 0.0505\ldots \approx 0.05 \tag{8}$$

Overlap values can potentially be used as measures of A's relevance given Q. However, a potential problem of using just the overlap to measure relevance between Q and A is that it gives value 0 in cases where A and Q are close but still do not overlap. However, because of the closeness of A and Q they might still have mutual relevance. For this reason we present next how *a closeness function c* will be used together with *overlaps* and *overlappedBy* values to calculate the relevance between Q and A.

The closeness function c between Q and A is a defuzzified value of the distance D. D is a fuzzy temporal interval and is calculated using an arithmetic operation fuzzy subtraction $\ominus$ (see Dubois and Prade, 1988, page 50), between a fuzzy query interval Q and a fuzzy annotation interval A, defined using the LR-notation (LR) as follows.

$$\begin{aligned} D = Q \ominus A \quad &= (Q_{begin} - A_{end}, \\ &\quad Q_{end} - A_{begin}, \\ &\quad \alpha_Q + \beta_A, \\ &\quad \beta_Q + \alpha_A)_{LR} \end{aligned} \tag{9}$$

We can then calculate the fuzzy extensions α and β for Q and A, respectively:

$$\begin{aligned} \alpha_Q &= Q_{begin} - Q_{fuzzybegin} &&= 89 \ B.C. - 222 \ B.C. &&= 133 \\ \beta_Q &= Q_{fuzzyend} - Q_{end} &&= 569 \ A.D. - 452 \ A.D. &&= 117 \\ \alpha_A &= A_{begin} - A_{fuzzybegin} &&= 510 \ B.C. - 490 \ B.C. &&= 20 \\ \beta_A &= A_{fuzzyend} - A_{end} &&= 89 \ B.C. - 222 \ B.C. &&= 133 \end{aligned} \tag{10}$$

and finally we get

$$
\begin{aligned}
D = Q \ominus A \ &= (89\ B.C. - 222\ B.C., \\
&\quad 452\ A.D. - 490\ B.C., \\
&\quad 133 + 133, \\
&\quad 117 + 20)_{LR} \\
&= (133, 942, 266, 137)_{LR}
\end{aligned}
\tag{11}
$$

The distance D in the above example between a fuzzy query interval Q and a fuzzy annotation interval A is hence a fuzzy temporal interval $D = (133, 942, 266, 137)_{LR}$.

To obtain the value for the *closeness* function c (the *closeness* measure) we follow the next steps.

1. Distance D is defuzzified to a crisp value d_{df} e.g. by calculating the Center of Area (COA) (see Zimmermann, 1996, pages 212—214) and taking the absolute value of it. Defuzzification is a procedure where a fuzzy set is transformed to a crisp number. By using COA we aim to take into account also the fuzzy extensions, i.e. where μ is in range $[0, 1)$, of the temporal intervals. COA is by definition the center of the area of the whole trapezoid, i.e. in our case the whole area between $T_{fuzzybegin}$ and $T_{fuzzyend}$. For example the defuzzification method Mean of Maxima (MOM) (Zimmermann, 1996) only considers the core of the fuzzy set where $\mu = 1$ i.e. in our case the area between T_{begin} and T_{end}. In our example, we get $d_{df} = 503.11$ for the defuzzified value of $D = (133, 942, 266, 137)_{LR}$.

2. The defuzzified distance $|d_{df}|$ is normalized to d_n, meaning its value is then between [0,1]. The normalization is done by dividing $|d_{df}|$ with the maximum d_{max} of all the defuzzified distances between all examined temporal interval pairs (in our case $d_{max} = 1111$). After this step, the closer d_n is to 0, the closer the two intervals are considered to be to each other.

3. The *closeness* measure c is calculated as $c = 1 - d_n$. As a result, c gets values close to 1 (i.e. better values[3]) when d_n is close to 0 (i.e. when intervals are close to each other).

The full equation for calculating closeness c from defuzzified distance d_{df} is thus $c = closeness = 1 - \dfrac{|d_{df}|}{|d_{max}|}$. In our example, we get for the value $c = 1 - \dfrac{503.11}{1111.11} \approx 0.55$.

---

[3]Values for overlap and overlappedBy are similarly considered better when close to 1.

We then combine the three measures (*closeness c*, *overlaps o_t* and *overlappedBy o_b*) to one relevance measure r that will get values in range [0,1] and provide a way to explicate weighting for them:

$$r = \frac{w_c * c + w_{ot} * o_t + w_{ob} * o_b}{w_c + w_{ot} + w_{ob}} \qquad (12)$$

where w_c, w_{ot} and w_{ob} refer to the weights used for the measures. In effect, the combined measure, r, is a weighted average of the three individual measures. The calculation of the weights is described in detail in Section 5.2.

3. The Case Study

3.1. Case Study: Ancient Milan

In the evaluation of the method we considered a dataset from the "Ancient Milan" project (Bandini et al., 2009). "Ancient Milan" aims at improving access to information concerning the archaeological heritage of Milan (Italy) through the utilization of Semantic Web and Web 2.0 techniques and tools. This data comes from heterogeneous data sources, which are integrated by means of the CIDOC CRM (Crofts et al., 2009) ontology.

Temporal information is provided mostly in the form of chronologies that are related to events that happened in ancient times (such as the the production of artifacts), or in modern or contemporary times (such as the phases of the research processes carried out by archaeologists and cultural heritage professionals). The case study takes into consideration the first kind of chronologies, the ancient times.

Here temporal intervals are defined by:

- references to "absolute chronology"[4], in terms of centuries and/or parts of centuries: e.g. the production of an artifact is attributed to sometime in the time span ranging from the end of the I century B.C. to the middle of the II century A.D.;

---

[4]A general distinction exists in Archaeology between absolute and relative chronologies. Absolute chronologies are based on a temporal scale of measurement, usually the calendar; therefore, events and periods are temporally qualified by measurement units, such as years, expressing their location on the calendar and their duration. On the contrary, relative chronologies are based on qualitative temporal relationships between events and periods, such as precedence, contemporaneity, succession, etc.

- references to historical periods: e.g. an artifact is attributed to the "Roman age".

3.1.1. About References to Absolute Chronology

References to absolute chronology are provided as highly consistent labels assuming one of the following forms:

- a single reference to a century: e.g. "I century B.C."

- a single reference to a part of a century: e.g. "middle I century B.C."

- a combination of these, e.g.:

  - "I century B.C. - IV century A.D."
  - "end I century B.C. - I century A.D."
  - "end III century A.D. - beginning IV century A.D."
  - "third quarter I century B.C. - VI century A.D."

At a more general level, each label in the dataset is the result of the combination of a few atomic elements, expressing temporal information with different degrees of imprecision, from parts of centuries to the era according to the Gregorian calendar. The possible combinations and values these elements can assume are expressed in the following set of specifications in the Backus-Naur Form (BNF):

> <label> ::= <reference> | <reference> "-" <reference>
>
> <reference> ::= <century> "century" <era> |
>   <part-of-century> <century> "century" <era>
>
> <part-of-century> ::= "first half" | "second half" |
>   "beginning" | "middle" | "end" |
>   "first quarter" | "second quarter" |
>   "third quarter" | "last quarter"
>
> <century> ::= "I" | "II" | "III" | "IV" | "V" | "VI"
>
> <era> ::= "BC" | "AD"

Figure 3 provides a schema that exemplifies different levels of imprecision denoting labels (X) with reference to the 1st century B.C. and the 1st century A.D. (Y).
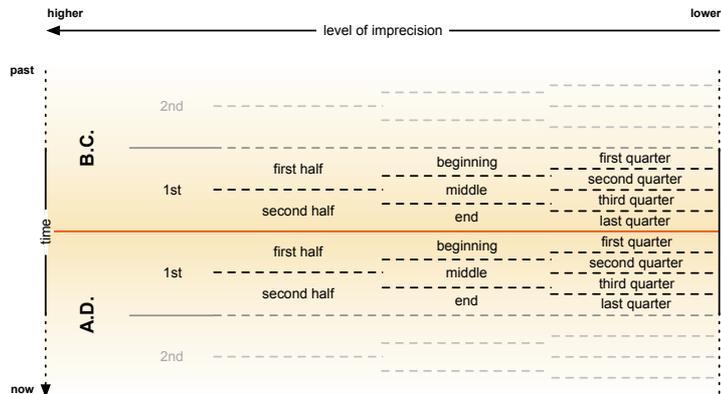
9

higher   level of imprecision   lower

past

B.C.

2nd

1st   first half   beginning   first quarter

second quarter

middle   third quarter

second half   end   last quarter

time

A.D.

1st   first half   beginning   first quarter

second quarter

middle   third quarter

second half   end   last quarter

2nd

now

Figure 3: A depiction of the possible references to absolute chronology showing different levels of imprecision.

3.1.2. About References to Historical Periods

Historical periods in the dataset of the case study are defined by imprecise temporal intervals. The link from the intervals to actual calendric dates can assume different forms: from the use of conventional dates (related to e.g. major historical events) to more generic indications (e.g. a whole century). The definition of the temporal boundaries of these intervals is made even more difficult by the fact that transitions from one historical period to another are generally continuous phenomena, and the duration of these transitions may vary a lot from case to case.

For example, the shift in the control of Milan from the Celts to the Romans happened quite gradually during more than a century, a period when elements belonging to both cultural spheres were present. Therefore, in order to define a flexible model for temporal intervals like the Celtic and the Roman periods a model based on fuzzy sets was adopted.

3.2. Determining Fuzzy Temporal Intervals from Label Information

The annotations in our case study dataset provided a solid base for determining fuzzy temporal intervals from the time labels using a combination of regular expressions and string parsing techniques.

In order to explicate the four temporal parameters needed for the representation of fuzzy temporal intervals as described in Section 2.1, the label was split into atomic elements. Figure 4 depicts examples of the guidelines

10

for creating fuzzy temporal intervals. These examples include e.g. "the beginning of a century", "the second half of a century", or "the last quarter of a century". For each type of atomic element its fuzziness was defined: e.g. the boundaries of a century were considered more fuzzy than boundaries of a smaller period such as the first quarter of a century.
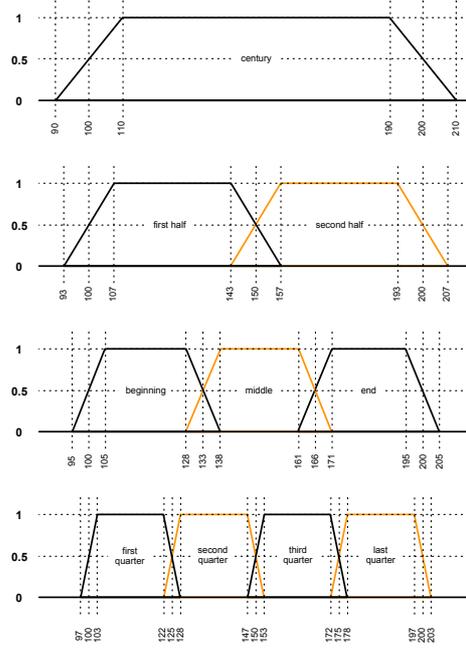

Figure 4: Guidelines for creating fuzzy temporal intervals.

To create these atomic elements the labels were checked whether they consisted of a single interval or two intervals (one representing the fuzzy begin interval and the other the fuzzy end interval). In the case of a single interval, the consistently formed label was processed in reverse order, i.e. from right to left. This procedure was due to the fact that the further right in the label string an atomic element is, the coarser is its granularity. For example, in the case of "first half II century A.D." the temporal element "A.D." determines the Gregorian era whereas the second element from the right, "II century", identifies a century within the defined era and the leftmost part, "first half", specifies a part of that century. As the elements are processed from right to left, the granularity gets finer. Finer granularity in our proposal aims to lead to more precise representations of the fuzzy boundaries.

11

If the label was formed of two intervals, both intervals were handled separately and were later combined. As a result the combined interval consisted of the fuzzified begin of the first interval and the fuzzified end of the second interval.

The resulting fuzzy temporal intervals were saved as RDF triples[5] where each of the four temporal properties (like $T_{fuzzybegin}$) describe the temporal instance.

4. The Evaluation

4.1. Evaluation Setting: Participants, Materials and Methods

In order to measure the correlation between different measures and human opinions we used the following evaluation setting. Together there were twelve human evaluators that were given the task to evaluate each of the query intervals according to each of the annotation intervals.

The query periods represent relevant historical phases in the ancient history of Milan, while the annotation periods refer to the temporal intervals when artifacts were produced according to archaeologists. This latter is generally called the "chronological attribution" of archaeological artifacts; in addition, the chronological attribution of structures and monuments has been integrated where available. Therefore, evaluators were asked to assess the relevancy of the chronological attributions of artifacts with respect to a given historical period; in more intuitive terms, this corresponds to evaluating the attribution of an artifact to a given historical period.

Eight of the evaluators were domain experts, with background from the fields of history, archaeology and museology; four evaluators were considered as average users. The choice of a mixed group of evaluators represents an interesting choice. On the one hand, expert evaluation offers the possibility of analyzing the results provided by the system by comparison to the results expected by professionals, and to fine-tune the retrieval mechanisms accordingly; on the other hand, evaluations by average users provide the possibility to understand how non-professionals perceive the interaction with the system and the relevance of its retrieval capabilities.

Twelve query intervals have been taken into consideration:

---

[5]http://www.w3.org/RDF/

- Six intervals belong to the periodization proposed by (Caporusso et al., 2007): the origins, the city of Insubres, from the Insubres to the Romans, between the end of the Roman Republic and the first centuries of the Roman Empire, the age of Maximianus, the christian city. This periodization offers the advantage of being specific to the city, and, at the same time, easily understandable with little background knowledge.

- Six intervals have been added by the domain expert, for their relevance in the context of the "Milano Antica" project: the pre-roman age, the late Republican period, the Roman age, Milan capital of the Roman Empire, the late Roman age, the Late Antique.

The selected query periods mostly refer to the Roman age, which represents the biggest and most significant phase in the ancient history of Milan. They show heterogeneous durations and degrees of temporal uncertainty; therefore, they provide an interesting and varied scenario for the evaluation of the system.

Evaluators rated annotation intervals with respect to query intervals with star ratings ranging from one to ten. Evaluators were instructed that all query/annotation pairs with no explicit rating will be treated as having zero stars.

4.2. An Interface for Enabling the Evaluation Setting

In order to efficiently evaluate the suitability of the calculated relevance measures, we created a Simile Timeline[6] representation of the historical and chronological periods of the case study data, see Figure 5. The interface is divided into two bands. The upper band shows the query periods, in the form of grey with light grey boundaries representing the fuzzy begin and fuzzy end. Users select one query period by clicking on its bar, and the lower band reconfigures itself in order to display the annotation periods. When users select an annotation period, the system displays the artifacts (on the right side of the interface) whose chronological attribution coincides with the selected annotation period. The selected query and annotation periods become highlighted; on the upper-right side, users can then attribute a star rating assessing the evaluation of the relevancy of the annotation period with respect to the query period. The evaluation happens in a highly interactive

---

[6]http://www.simile-widgets.org/timeline/

and visual way. The average time employed for evaluating all the possible combinations of the query and annotation periods (nearly 800 ratings) was four hours i.e. 18 seconds per rating.
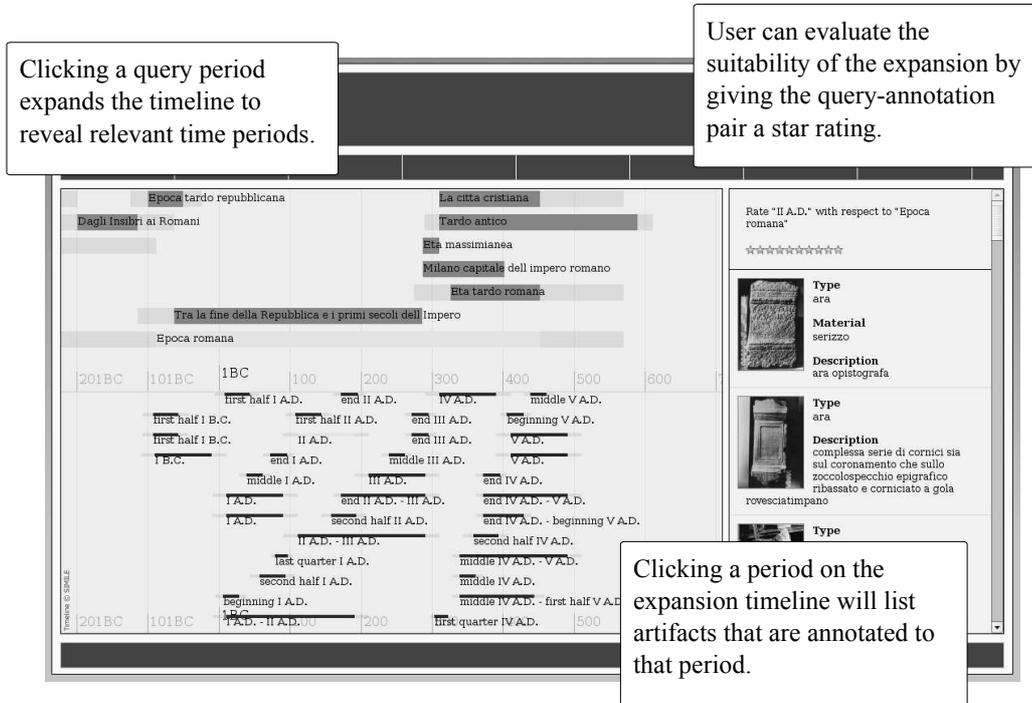

Figure 5: Interface to evaluate relevance measures.

5. Results

5.1. Ratings from Evaluators

Twelve evaluators $(E_1 \ldots E_{12})$ evaluated all 12 query intervals with respect to all 66 unique annotation intervals as described in the previous section. In other words, each query/annotation pair was rated by each of the same twelve evaluators. Recall that all query/annotation pairs with no explicit rating were treated as having zero stars. The user agreement about relevance levels between queries and annotations was analyzed using weighted kappa (Cohen, 1968). Weighted kappa coefficients are commonly used to quantify inter-rater reliability when ordinal categories are used in rating.

14

Weighted kappa assigns less weight to agreement if ordinal categories are further apart. For example, a disagreement of eight stars versus nine stars in rating would still be credited with partial agreement, whereas a disagreement of ten versus zero stars would be considered as no agreement at all. A weighted kappa coefficient of 1.0 means maximum possible agreement and a weighted kappa coefficient of 0.0 means the lack of agreement.

In our case the average of weighted kappa's calculated pairwise was 0.83, ranging from 0.64 up to 0.94. The closer examination of pairwise kappa's revealed that among twelve evaluators one evaluator gave notably different relevance assessments than others. In other words, her agreement with other evaluators was lower compared to other evaluators' agreement with each other. However, because all eleven other evaluators agreed more about the relevance assessments the average of the kappa's was 0.83. Values between 0.81-1.0 are considered as a very good agreement (Altman, 1991). Because of this level of user agreement we decided to use the average of the user opinions as the gold standard for weighting the relevance measures, and for comparing the final results with. These will be discussed in the next sections.

5.2. Weights for a Combined Relevance Measure

In order to calculate how each of the relevance measures (calculated using the methods described in Section 2.1) should be weighted we applied multiple linear regression to obtain regression coefficients for each relevance measure using the averaged evaluator opinion as response observations. More precisely, we ran linear regression to obtain weights for different combinations (e.g. all three measures, or just *overlaps* and *closeness*, or just *overlap* and *overlappedBy*, and so on). Precision and recall analysis (see the next section for an outline of it) showed that a measure that combines the weighted *overlaps* and *closeness* measures together (without *overlappedBy*) performs best. Linear regression gave the values $w_c = 0.13$ (*closeness*) and $w_o = 0.73$ (*overlap*) to be further used to weight measures as described in Section 2.1. Because the combination that used also *overlappedBy* measure did not enhance the results, we will use $w_{ob} = 0$ in this setting. In other words, we tried also a measure that combines all three weighted measures together, including *overlappedBy*. However, it gave about the same results as if we used the combination of only two weighted measures, *overlaps* and *closeness* together. Thus we used Occam's razor and did not include *overlappedBy* as a part of the combined measure.
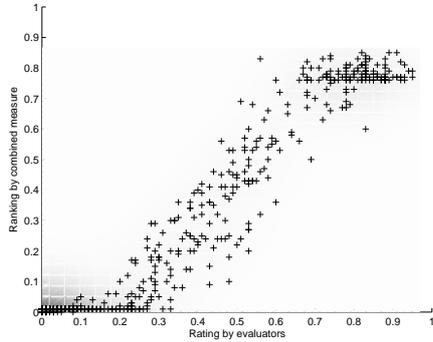
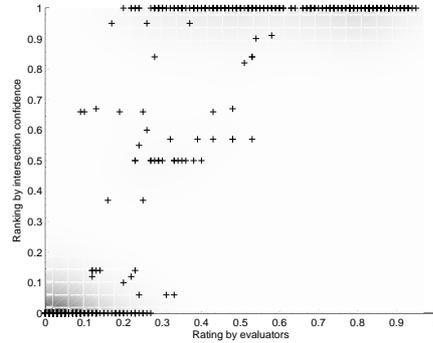Figure 6: Scatter plot of relevance ranking based on combination of weighted measures.

Figure 7: Scatter plot of relevance ranking based on intersection confidence.

The correlation between the average of the evaluator ratings (x-axis) and the combination of these two weighted measures (y-axis) is visualized as a scatter plot graph in Figure 6. This reveals that the correlation between the combined measure and the evaluator ratings is strong. Figure 7 shows a correlation graph between the evaluator ratings and the intersection confidence (Nagypál and Motik, 2003) for a comparison. The intersection confidence evaluates the degree (in range [0,1]) to which the temporal relationship called "intersects" exists between two fuzzy temporal intervals. More precisely, the intersection confidence expresses the confidence that $A \cap Q \neq \emptyset$, i.e. the confidence that the intersection of A and Q is not empty. Intersection confidence was calculated using Equation (13).

$$Intersection \ \ confidence = sup \ \ min(A, Q) \tag{13}$$

where $min(A, Q)$ is the fuzzy intersection of the A and Q, and $sup \ \ min(A, Q)$ (supremum) is the maximum confidence of membership of any time point in the intersection.

From Figures 6 and 7 it can be clearly seen that the combination of weighted measures of Figure 6 has more linear correlation with the users' opinions than that of intersection confidence of Figure 7.

5.3. Precision and Recall Analyzed

Precision figures of standard 11 recall levels were calculated in order to evaluate quantitatively the performance of different measures. First of all, we compared the performance of individual measures *overlaps*, *overlappedBy*

16

and *closeness*, and the combined measure (combination of *overlaps* and *closeness*). We also calculated the performance of the intersection confidence (Nagypál and Motik, 2003). As a baseline we used a binary overlap measure between the crisp intervals: if crisp areas of intervals overlap then the value will be 1 (relevant) and if not then the value is 0 (non-relevant).

The averaged user ratings were considered as a golden standard. Ratings in range 1–10 stars were considered as relevant and annotations that got 0 stars (after rounding the averaged user opinions to the closest star rating) were considered as non-relevant for a given query. Figure 8 depicts the resulting precision-recall curve. The combined measure clearly performed best, closeness measure was the second, and *overlaps* and *overlappedBy* competed for the third place. However, *overlaps* is mostly just above the *overlappedBy*. The intersection confidence measure and the baseline got the lowest values.
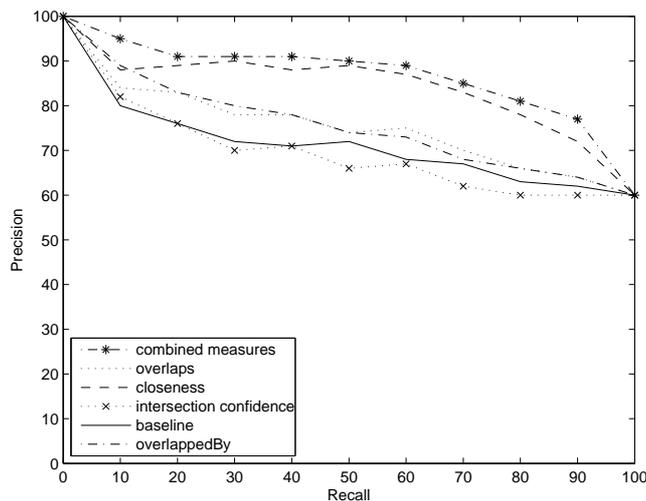

Figure 8: Average recall versus precision figures for each measures used to rank the results.

The standard calculation of precision and recall evaluates the performance of methods based on binary relevant vs. non-relevant values. However, the evaluation in our case contained also multiple grade relevance assessments in range 0 to 10 stars. For this reason we also report the results of using generalized precision and recall (Kekäläinen and Järvelin, 2002) that is intended to reward methods retrieving highly relevant documents. The generalized

17

precision and recall applied to our case are defined as follows. Let R be the set n annotation periods from the knowledge base $K = \{a_1, a_2, \ldots, a_n\}$ in response to a query Q, $R \subseteq K$. Let the annotation periods a_i in the knowledge base have relevance scores $r(a_i)$ in range [0,1]. Generalized precision is then computed using the Equation (14) and generalized recall using the Equation (15).

$$Generalized\ \ precision = \sum_{a \in R} r(a)/n \qquad (14)$$

$$Generalized\ \ recall = \sum_{a \in R} r(a) / \sum_{a \in K} r(a) \qquad (15)$$

Figure 9 depicts the resulting generalized precision-recall curve. Here, the combined measure had the highest values, *overlaps* was the second, the intersection confidence as the third, and then *closeness*. The baseline and *overlappedBy* have the lowest values. Note that generalized precision-recall curves often fall a lot lower than traditional precision-recall curves. Finally, the differences between rankings given by each type of measure were found out to be statistically significant using the Friedman test (p=0). The Friedman test is a non-parametric statistical test intended to determine if differences between rankings are significant (Friedman, 1937). The Friedman test makes no assumptions about how the data is distributed (e.g. if the data is normally distributed or not), and hence it was chosen in our case. A post-hoc analysis was run as a pairwise Friedman test between all combinations of rankings in order to ensure the statistical significance of differences. The analysis showed the statistical significance (p=0) of pairwise differences of rankings.

6. Discussion

The results of comparing the rankings given by the method and the ratings given by the evaluators seem promising as there is an apparent correlation between the rankings and the ratings. The precision and recall curve shows that the new combined measure performed best. Among single measures the closeness performed well in traditional precision and recall analysis. This might be due to the fact that other measures measure the level of overlap (or confidence) in different ways and simply do not notice, if intervals are quite close, but do not overlap.
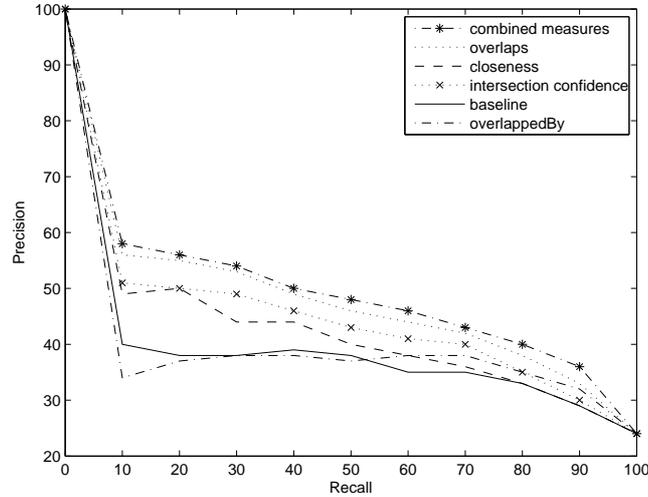
Figure 9: Generalized precision-recall curve depicting the results.

The generalized precision and recall curve shows that the combined measure is again the best. However, this time the overlaps measure is the second. So it seems that of the single measures, the overlaps measure is very important; the more the annotation interval was immersed within the query interval, the better rating evaluators gave for this query-annotation pair. All in all best results were obtained by combining weighted overlaps and closeness measures together.

However, it should be noted that the scope in our case study was narrow: the case study concerned a restricted spatial and cultural area. Moreover, the case study took into consideration only the chronological information which the references to intervals carried, without addressing the characteristics and value of cultural attributions.

Some periods are also open to different possible degrees of subjective temporal characterization, reflecting e.g. different scientific opinions. For example, the "Late Roman age", while evidently ending together with the end of the Roman age (which in turn is defined by a fuzzy end), can be differently characterized with respect to its beginning, on the basis of e.g. different scholarly opinions on the meaning of "late".

From the modeling point of view the presented approach is generalizable

19

to many different domains (in addition to the cultural heritage domain) and scales (e.g. "the beginning of the work week" vs. "the glacial age"). In essence, fuzzy temporal intervals are intended for modeling imprecision of temporal expressions, independent of the scale or the theme. Same applies to the proposed relationships: *overlaps*, *overlappedBy*, and *closeness* may be determined between any two fuzzy temporal intervals in a similar manner as was shown for query and annotation periods. The weighted combination of *overlaps*, *overlappedBy*, and *closeness* is also generalizable to different domains. However, the weights might vary in different domains (e.g. personal planning vs. history of natural sciences). The presented procedure to assign weights through a user evaluation and linear regression resulted in a combined measure that provided best results in terms of precision and recall in our cultural heritage case study. Hence we expect that a similar kind of procedure could be used to reveal weights for temporal relationships for information retrieval tasks in other domains of interest. However, this should of course be confirmed by a future study. For example, one research question could be to study if evaluators consider the usefulness of overlap and closeness of time periods in information retrieval domain- and scale independently.

7. Related Work

The general properties of time ontologies have been explored e.g. in Vila (1994). There have been discussions e.g. whether the basic primitive is the interval (period) or the point. Other properties for time are characterized by whether time is discrete or dense, bounded or unbounded, and what type of precedence the time ontology allows: linear, branching, parallel or circular (cyclic). Allen (1983) has presented a set of the 13 primitive interval relations that exclusively correspond to every possible simple qualitative relationship that may exist between a pair of crisp intervals.

Visser (2004) has identified the following different types of boundaries for periods: 1) exact boundaries, 2) persistent boundaries, 3) unknown boundaries and 4) fuzzy boundaries. Our interest was in 1) and 4) type of period boundaries: the presented method can be used to find the annotation intervals that are relevant to a query interval.

Several projects dealing with cultural heritage and the web present methods for the representation of temporal information and tools for retrieving it, see e.g. Hyvönen et al. (2009) and Schreiber et al. (2006). Johnson (2008) reviews the approach adopted in the TimeMap-ECAI project and provides

20

an updated set of proposals and advancements towards the integration of Web 2.0 techniques and tools. In particular, he introduces some considerations about new modalities for developing interactive timelines concerning historical events.

Supporting chronological reasoning in Archaeology is the focus of Doerr et al. (2004), where a formal framework is introduced, which is the core of the temporal representation in the CIDOC CRM (Crofts et al., 2009). This framework relies on the idea of "events as meetings", and discusses different and relevant issues such as temporal indeterminacy; however, being a theoretical and high-level work, it does not enter into the details of the actual use of the theory in real application scenarios (which can be found in the documentation of projects using the CIDOC CRM). There is also earlier work (Accary–Barbier and Calabretto, 2008) that describes a specific case in the field of digital libraries dealing with the possibility of comparing different temporal models of knowledge in archaeological documentation, that may emerge from fuzzy or even contrasting chronologies.

Schockaert and Cock (2008) examined problems related to reasoning about qualitative and metric temporal relations between fuzzy time period. Schockaert et al. (2008) discusses about how to measure e.g. the degree to which "the beginning of a fuzzy temporal interval A" *is long before* "the beginning of a fuzzy temporal interval B". Their approach uses fuzzy orderings between time points (e.g. between the beginning of A and the beginning of B) to define temporal relations. The difference is that in our approach we aimed at measuring the closeness of the whole A to the whole B using the fuzzy subtraction. Moreover, the generic goal in Schockaert et al. (2008) was to provide a fuzzification of Allen's temporal interval relations. In our approch the goal was to determine the relevance between imprecise temporal intervals, and to provide a combined measure for ranking results according to a query. While Schockaert et al. (2008) discussed the usefulness of their approach within the context of question answering systems our context was cultural heritage information retrieval. Another approach, provided by Ohlbach (2004), can also be used for expressing imprecise temporal relations. However, they do not provide any handling of closeness of two fuzzy temporal intervals.

Nagypál and Motik (2003) introduced a mechanism to evaluate whether e.g. a crisp temporal relationship called *intersects* holds between two fuzzy temporal intervals. The result is a value explicating the level of this confidence. We evaluated its usability also for relevance calculation. However, as

we showed the combined measure (overlaps and closeness together) was better in terms of precision and recall. Visser (2004) proposed to calculate the overlap between two fuzzy temporal intervals, but did not provide any evaluation results conforming the usability of the overlap relation. Also, closeness was neither considered nor tested in their study.

8. Conclusions

The proposal and experiments presented in this paper provide an approach for matching time intervals. Imprecision of temporal intervals was modeled using fuzzy sets and a method was developed in order to obtain a better match between human and machine interpretations of periods in information retrieval. For this we proposed to calculate overlappings and closenesses between annotation and query intervals, and showed how they can be combined together. The models and the method were tested in a real environment with data from the archaeology domain. Twelve evaluators rated each of the annotation intervals according to each query interval. These results were used as a basis for analyzing the correlation between the ranking given by the method and the ratings given by the human evaluators using linear regression. The results of the linear regression were further used as weights to fine tune the method. Both the standard precision and recall test and the generalized precision and recall test showed that the combination of overlaps and closeness measures (the combined measure) performed better than any of the single measures alone. The method could be used in e.g. suggesting items from approximately the same period as the reference query period and also for ranking the relevance of more distant periods of time.

---

[7]http://smartmuseum.eu/

and Technological Development and in the National Semantic Web Ontology Project in Finland[8] (FinnONTO) 2003–2007, 2008-2010 funded mainly by the Finnish Funding Agency for Technology and Innovation (Tekes). We gratefully acknowledge the support by Regione Lombardia, the SIRBeC team and the SIRBeC partners in providing the dataset for this work.

References

Accary–Barbier, T., Calabretto, S., October 2008. Building and Using Temporal Knowledge in Archaeological Documentation. Journal of Intelligent Information Systems 31 (2), pp. 147–159.

Allen, J. F., 1983. Maintaining knowledge about temporal intervals. Communications of the ACM 26 (11), pp. 832–843.

Altman, D. G., 1991. Practical Statistics for Medical Research. Chapman & Hall.

Bandini, S., Locatelli, M., Mantegari, G., Simone, C., Vizzari, G., 4–6 November 2009. An integrated approach to the valorization of cultural heritage. In: Proceeding of the 2009 AICA Conference. Rome.

Caporusso, D., Donati, M. T., Masseroli, S., Tibiletti, T., 2007. Immagini di Mediolanum. Civiche Raccolte Archeologiche e Numismatiche di Milano, Milano.

Cohen, J., 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin.

Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M., March 2009. Definition of the CIDOC Conceptual Reference Model Version 5.0.1. Tech. rep., ICOM/CIDOC CRM Special Interest Group.

Doerr, M., Plexousakis, D., Kopaka, K., soula Bekiari, C., 2004. Supporting Chronological Reasoning in Archaeology. In: Proceedings of Computer Applications and Quantitative Methods in Archaeology CAA'04. Prato, Italy.

---

[8]http://www.seco.tkk.fi/projects/finnonto/

Dubois, D., Prade, H., 1988. Possibility Theory: An Approach to Computerized Processing of Uncertainty. Plenum Press, New York.

Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32 (200), pp. 675–701.

Hyvönen, E., Mäkelä, E., Kauppinen, T., Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K., Kuittinen, H., Viljanen, K., Tuominen, J., Palonen, T., Frosterus, M., Sinkkilä, R., Paakkarinen, P., Laitio, J., Nyberg, K., 2009. CultureSampo—Finnish culture on the Semantic Web 2.0. Thematic perspectives for the end-user. In: Proceedings, Museums and the Web 2009, Indianapolis, USA.

Johnson, I., 2008. Mapping the fourth dimension: a ten year retrospective. Archeologia e Calcolatori 19, pp. 31–43.

Kekäläinen, J., Järvelin, K., 2002. Using graded relevance assessments in IR evaluation. Journal of the American Society for Information Science and Technology 53 (13), pp. 1120–1129.

Nagypál, G., Motik, B., 2003. A fuzzy model for representing uncertain, subjective, and vague temporal knowledge in ontologies. In: CoopIS/DOA/ODBASE. pp. 906–923.

Ohlbach, H. J., 2004. Relations between fuzzy time intervals. In: Proceedings of the 11th International Symposium on Temporal Representation and Reasoning (TIME 2004). Tatihou Island, Normandie, France, pp. 44–51.

Schockaert, S., Cock, M. D., 2008. Temporal reasoning about fuzzy intervals. Artificial Intelligence 172 (8-9), pp. 1158–1193.

Schockaert, S., Cock, M. D., Kerre, E. E., 2008. Fuzzifying allen's temporal interval relations. IEEE Transactions on Fuzzy Systems 16 (2), pp. 517–533.

Schreiber, G., Amin, A., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Hollink, L., Huang, Z., van Kersen, J., de Niet, M., Omelayenko, B., van Ossenbruggen, J., Siebes, R., Taekema, J., Wielemaker,

J., Wielinga, B., 2006. MultimediaN E-Culture demonstrator. In: Proceedings of the Fifth International Semantic Web Conference (ISWC'06). No. 4273 in Lecture Notes in Computer Science. USA, pp. 951–958.

Vila, L., 1994. A survey on temporal reasoning in artificial intelligence. AI Communications 7 (1), pp. 4–28.

Visser, U., 2004. Intelligent information integration for the Semantic Web. Springer-Verlag, Berlin Heidelberg, New York.

Zadeh, L. A., 1965. Fuzzy sets. Information and Control 8 (3), pp. 338–353.

Zimmermann, H.-J., 1996. Fuzzy Set Theory and its Applications, 3rd Edition. Kluwer, Dordrecht.