Representing and Reasoning with Intervals

Erica Yuan-Ting Huang[1], Clinton Smyth[1], David Poole[2]

[1]Georeference Online Ltd., Vancouver, B.C., Canada
[2]Department of Computer Science, University of British Columbia
E-mail: ehuang@georeferenceonline.com

1. Abstract

Geologists around the world often specify numerically measured information such as age or speed with non-numerical words such as "Jurassic" or "fast" rather than actual numerical values or ranges of numerical values. These qualitative descriptions, which may adhere to different standards, can be applied at different levels of detail or abstraction. This presents difficulties in building knowledge-based systems that have to compare temporally and spatially related information such as geological age or landslide movement rates.

This paper discusses the representation of intervals and the reasoning required when comparing interval information. It proposes a format, which can work with different interval domains, for representing intervals based on a current software system developed for qualitative probabilistic matching. This is a necessary first step in being able to reason with various interval specification standards that have evolved in different parts of the world.

Not only can this framework be used in the geological domain, such as mineral exploration and geological hazard evaluation, but it can also be used for intervallic reasoning in other scientific disciplines.

2. Introduction

A software system for matching semantic network descriptions of instances (e.g. a particular mineral deposit) and models (e.g. a type of mineral deposit) has been developed using a qualitative probability approach (Smyth and Poole, 2004). The commercial implementation of this system is called MineMatch® and deals with mineral deposits. The system is specifically designed to reason correctly when making comparisons between deposits, described at various levels and abstractions, with hierarchical classifications such as the taxonomy of rocks (BGS, 1999). Taxonomic reasoning can be used, to a limited extent, to reason with hierarchical representations of interval attributes, such as geological age or landslide movement rates. However, it limits flexibility in describing ranges such as "from Jurassic to middle Triassic" or "from slow to medium speed". It is also not sensitive to overlapping intervals, or to the degree of separation between non-overlapping intervals. Hence special representations and reasoning algorithms are necessary for correctly working with interval data.

3. Representing Intervals

Before computer systems can be used to match instance and model descriptions of natural phenomena, they require two types of input data. One is a knowledge base, or an ontology (Gruber, 1993), that provides the vocabulary used to describe the natural phenomena. The other type of input data is the actual descriptions of instances or models of natural phenomena in the real world.

To build ontologies for the current system, we, like OWL, the Web Ontology Language (Patel-Schneider et. al, 2004), provide a special data construct for hierarchical classification vocabulary, such as the rock classification system, mineral types, element types, or, what we focus on in this paper, interval attributes like the geological time scale.

However, instead of considering the term for an interval attribute, which is a linguistic reference to an interval, as a class in a tree structure, it is more accurate to think of it as a range on a number line. The value "middle Jurassic", for example, should be defined as a sub-range of "Jurassic" instead of being a sub-class of "Jurassic".

Interval terms, as we define them, encompass the start and end values of the interval and all the values in between. These values, along with their units, should also be defined in the knowledge base. Therefore, the system requires the ability to represent a pair of values as a single value as shown in Figure 1.

<Early Jurassic, subRangeOf, Jurassic>
<Middle Jurassic, subRangeOf, Jurassic>
<Early Jurassic, containsValue, [-206 ma, -178 ma]>
<Jurassic, containsValue, [-206 ma, -144 ma]>

Figure 1. Left: Data representation of interval information in the system; Right: Graphical representation corresponding to the information described by triples on the left (ma = millions of years).

Our system requires that, at the highest interval resolution made available to the user, the following conditions are met by the terms used to represent these intervals: (1) They must cover the entire range which needs to be referenced by the domain of interest; (2) No overlaps or gaps are allowed. Examples of these intervals are Middle and Late Jurassic in Figure 1.

Lower resolution terms may be defined which bracket together contiguous intervals. Their boundaries are restricted to occurring only where boundaries exist between the highest resolution intervals. An example of such an interval is Jurassic in Figure 1.

Once a vocabulary is defined in the ontology, it can be used to describe real world phenomena. OWL uses a three-tuple structure <Object, Property, Value> to describe things in the world and relations between these things. We extend the OWL schema into a five-tuple structure <Object, Property, Value, Frequency, Reference> for representing snippets of information. The frequency parameter contains a value "present" or "absent" for instance descriptions and a value from the five-value scale (always, usually,

sometimes, rarely, never) for model descriptions. Such additions enrich the knowledge base and allow the system to deal with uncertainty in model descriptions. The reference parameter allows the user to look up the source of the input data, which is an essential component for any knowledge interchange framework, such as the semantic web (Berners-Lee, 1998). An example of an instance description reads: <mineral assemblage A, containsElement, Cu, present, Smith 2004>, and an example of a model description reads: <mineral assemblage, containsElement, Cu, usually, Smith 2004>.

For flexibility in describing an interval, the system must allow reference to a pair of linguistic terms. For example: <limestone X, isOfGeologicalAge, [Carboniferous, Early Permian], usually, Smith 2004>. If the user only needs to specify a range using a single word, such as "Jurassic", the system will automatically store this information as a pair [Jurassic, Jurassic] to keep format consistency.

This architecture allows a user to describe interval data with familiar terms, where the necessary numeric boundaries can be referenced in the system but not explicitly seen in the descriptions.

4. Reasoning with Intervals

Two key factors in conducting interval comparisons are the relative sizes of the compared intervals and their relative positions on the number line.

Inspired by Allen's interval algebra (Allen, 1983), five categories are used to describe possible relations between two intervals in our system: exact range, sub-range, super-range, overlapping ranges, and non-overlapping ranges. The first three of these appear in the "During/Include" column in Figure 2 below. In addition to depending on the position relationships documented by Allen, the degree of match also depends on the relative size of the compared intervals.

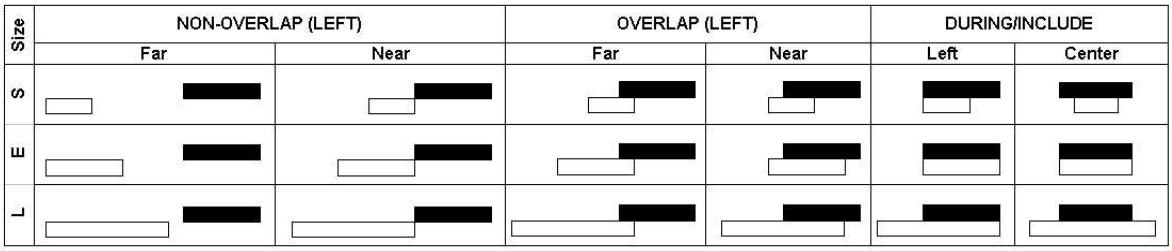


Figure 2. An illustration of different interval relations with different magnitudes. The "Size" column specifies the size of the "Target" interval (white rectangle): S = small, E = equal, and L = large. The Target interval is an interval (possibly one of many) against which the "In-Focus" interval of interest is compared (black). This table only summarizes half of the interval relations. Equivalent relations between Target and In-Focus intervals, but with Target intervals displaced to the right, have not been shown

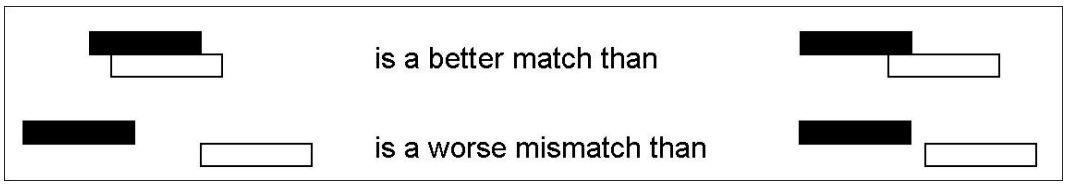


Figure 3. Degree of matches and mismatches we want to distinguish when comparing overlapping and non-overlapping intervals.

Representing the intervals as ranges on the number line instead of classes in a tree structure allows the system to perform matching on overlapping ranges and non-overlapping ranges, as illustrated in Figure 3 above. The interval boundary values captured when defining ranges in the ontology make possible similarity measures base on the "numerical" amount of overlap, or distance between compared intervals. However, we do not currently compare intervals in this way, but rather treat each single interval measure (of a time period, a landslide movement rate, a slope category) as a concept, and compute the similarity measure using "semantic distances" along the number line. The validity of this approach relies on the interval-defining terms having start and end values which are relevant to the domain in which they are being used. This is valid whenever the importance of number-line-related differences is better reflected by the linguistic terms than by the absolute differences in numerical values.

To match intervals that have some overlap is quite straightforward; we can use the proportion of overlap as our estimate of the level of match. However, to be able to distinguish between close matches and not-so-close matches when comparing non-overlapping ranges, discrete probability models of the ranges are built, by exponentially dropping down the probability at each semantic step, to approximate the probability of a value occurring within a particular range not specified by the user. The actual values of the match scores do not reflect actual probability of one interval being the same as the other; they are similarity measures used for ranking how good the match is and are meaningful in a relative sense.

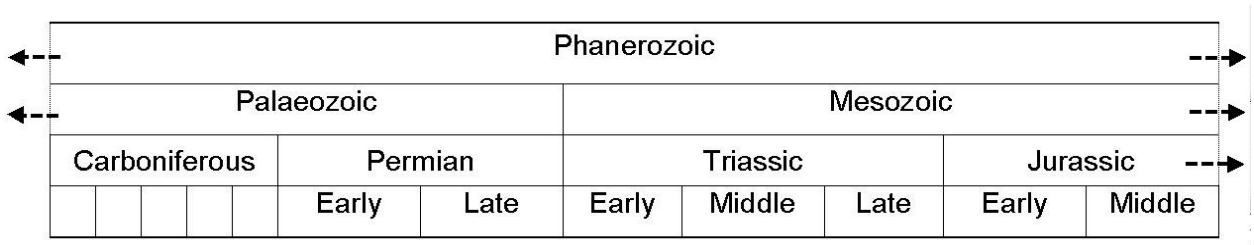
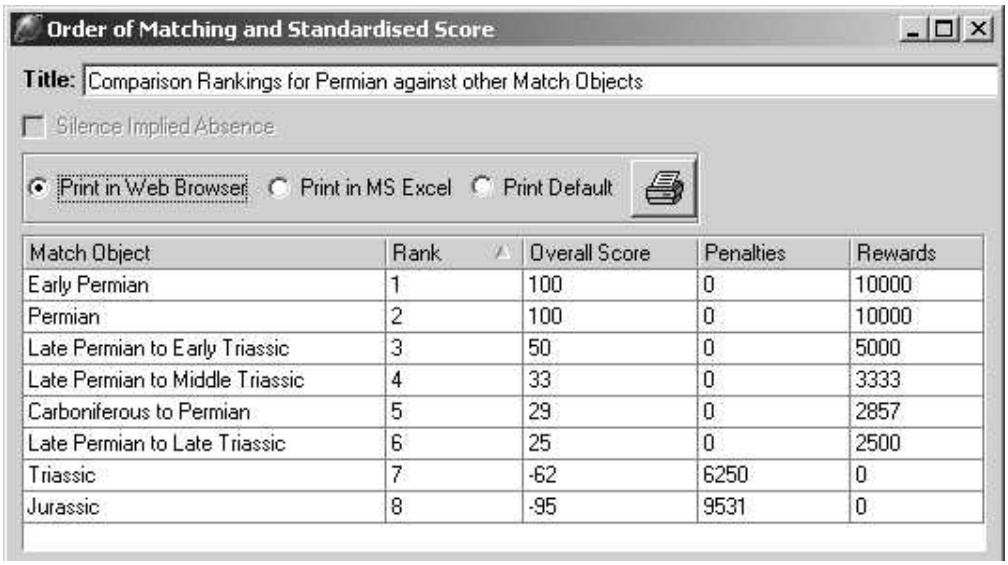


Figure 4. Part of the geological timeline.



Match Object	Rank	Overall Score	Penalties	Rewards
Early Permian	1	100	0	10000
Permian	2	100	0	10000
Late Permian to Early Triassic	3	50	0	5000
Late Permian to Middle Triassic	4	33	0	3333
Carboniferous to Permian	5	29	0	2857
Late Permian to Late Triassic	6	25	0	2500
Triassic	7	-62	6250	0
Jurassic	8	-95	9531	0

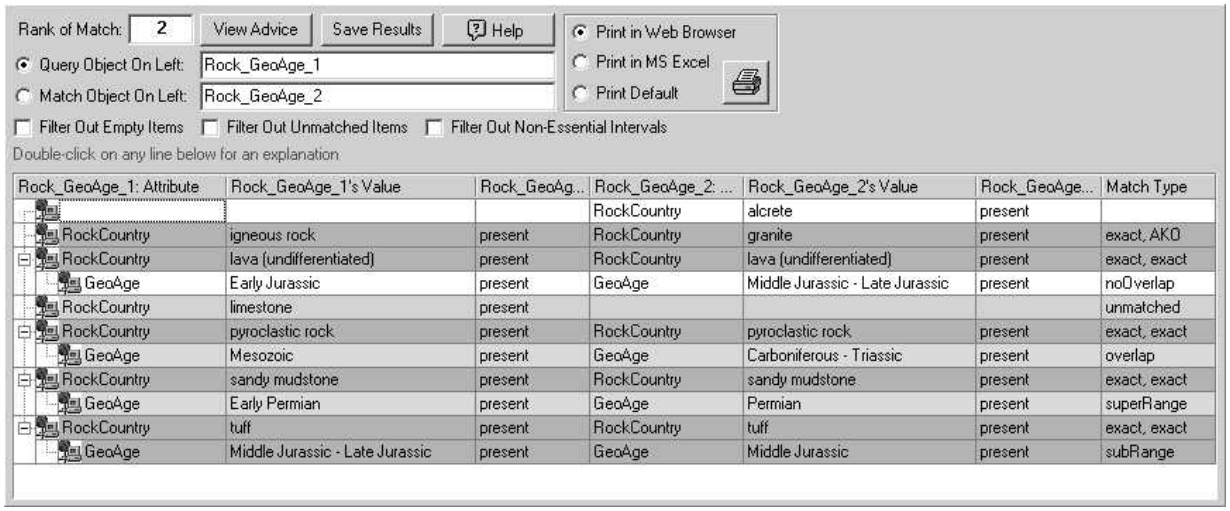
Figure 5. Sample results for comparing the range "Permian" with some test age ranges.

The reasoning becomes more complicated when users would like to assign different frequency levels for more than one adjacent interval in a model description. For example, a user may wish to say "landslide type X usually moves very fast, sometimes extremely fast." Nevertheless, the principle of building a discrete probability model to represent this interval information for matching purposes is the same.

5. Example Applications

There are many fields in geology that require the specification and comparison of interval data, such as mineral exploration (age of rocks, concentration in rock composition, grain size etc.) and landslide hazard evaluation (landslide speed, slope etc.).

Figure 6 below illustrates various matches and mismatches as they are recognized and reported for a comparison of rock descriptions produced by the PlutonMatch system (Smyth, 2005).



Rock_GeoAge_1: Attribute	Rock_GeoAge_1's Value	Rock_GeoAg...	Rock_GeoAge_2: ...	Rock_GeoAge_2's Value	Rock_GeoAge...	Match Type
			RockCountry	alcrete	present	
RockCountry	igneous rock	present	RockCountry	granite	present	exact, AKO
RockCountry	lava (undifferentiated)	present	RockCountry	lava (undifferentiated)	present	exact, exact
GeoAge	Early Jurassic	present	GeoAge	Middle Jurassic - Late Jurassic	present	noOverlap
RockCountry	limestone	present				unmatched
RockCountry	pyroclastic rock	present	RockCountry	pyroclastic rock	present	exact, exact
GeoAge	Mesozoic	present	GeoAge	Carboniferous - Triassic	present	overlap
RockCountry	sandy mudstone	present	RockCountry	sandy mudstone	present	exact, exact
GeoAge	Early Permian	present	GeoAge	Permian	present	superRange
RockCountry	tuff	present	RockCountry	tuff	present	exact, exact
GeoAge	Middle Jurassic - Late Jurassic	present	GeoAge	Middle Jurassic	present	subRange

Figure 6. Result of comparing geological age of rocks. The first three columns list attribute, value, and presence/absence status of information about the In-Focus instance, the next three columns list attribute, value, and presence/absence status of information about the Target instance, and the last column list the match types. Exact and AKO (a-kind-of) are match types recognized using taxonomic reasoning. Sub-range, super-range, overlap, and "no overlap" are match types recognized during interval reasoning.

6. Conclusion

A new approach of reasoning with interval data has been developed, which facilitates the similarity ranking of instance and model descriptions. This approach has broad application, and can greatly simplify the interface between humans and computer systems developed to reason in a human understandable manner.

7. Reference

Allen, J.F., 1983: Maintaining Knowledge about Temporal Intervals, Communications of the ACM, 26(11), 832-843.

Berners-Lee, T., 1998: What the Semantic Web Can Represent, World Wide Web Consortium (W3C). URL: http://www.w3.org/DesignIssues/RDFnot.html

British Geological Survey (BGS), 1999: Rock Classification Scheme. URL: http://www.bgs.ac.uk/bgsrcs/home.html

Gruber, T. R., 1993: A Translation Approach to Portable Ontology Specifications, Knowledge Acquisition, 5(2), 199-220. URL: http://www-ksl.stanford.edu/knowledge-sharing/papers/README.html#ontolingua-intro

Patel-Schneider P.F., Hayes P., and Horrocks I., 2004: OWL Web Ontology Language Semantics and Abstract Syntax, W3C. URL: http://www.w3.org/TR/owl-semantics/

Smyth, C. and Poole, D., 2004: Qualitative Probabilistic Matching with Hierarchical Descriptions, KR-04, Whistler, B.C., Canada.

Smyth, C. 2005: PlutonMatch: A computer system for describing and comparing plutons. URL: http://www.georeferenceonline.com/plutonmatch.