Qeios

Peer Review

Review of: "Evidence Admissibility as a Requester- and Context-Relative Retrieval Constraint: A Formal Specification and Empirical Characterization"

Iliano Fasolino[1]

1. Department of Computer Science, University of Milan, Italy

This version does what a revision should do. On v1, John McCrae asked whether Eligibility Before Ranking was more than metadata pre-filtering, and v2 answers with an experiment: configuration I, a flat filter over the same materialized closure facts, lands at a RIER difference of +0.0007 (95% CI [+0.0000, +0.0021]), and the architectural claim and the Universal Admissibility Ontology label are both withdrawn (Sections 9.2 and Notes). The ontology is now OWL 2 DL, checked with HermiT against thirteen competency questions, and Section 6.6 records the defect that inflated the flat baseline's relational catch rate from 0.576 to 0.946. The tables agree with each other: for relevance-only retrieval, AP@5 0.2536 plus RIER 0.2121 recovers the Precision@5 of 0.466 in Section 7.1. My own work is on corrupted retrieved context, so I read admissibility gating as the defensive side of that problem.

Marek Pawlicki and Haya El Ghalayini have covered length and structure (she also asks for a methodology diagram), and Sietse Wieringa asked whether failing one dimension should make an artifact wholly inadmissible. None of these v2 reviews has an author reply yet, and I will not repeat them.

On novelty, I read the paper as the authors do: Section 2 calls the contribution "a composition rather than a discovery," and Section 9.3 restates it. The closed-world/open-world asymmetry of Section 5 is what the open-world assumption gives any OWL ontology without closure axioms, so I take it as a clearly stated consequence rather than a new result.

The ternary form is the stated difference from ContextNest (Section 2.2), yet as written, the requester never appears as an argument. Section 4.1 reifies AdmissibilityDecision with functional properties

 doi.org/10.32388/8MN61Y

aboutArtifact and inContext only, while AdmissibleArtifact is defined as the complement of InadmissibleArtifact within EvidenceArtifact, and the two are declared disjoint, so the verdict classes sit on artifacts (the Section 4.3 control fixture asserts an artifact to be AdmissibleArtifact). If facts for two contexts were loaded into one knowledge base, an artifact entailed admissible in one and inadmissible in the other would be an individual in two disjoint classes, and the knowledge base would be inconsistent, which conflicts with Definition 1. Either each context is reasoned over separately, in which case the context- and requester-relativity of Definition 1 lives in the materializer rather than in the ontology, or the verdict classes should range over AdmissibilityDecision. Please say which, and show the axioms that connect AdmissibilityDecision to the two artifact classes. Empirically, each query/context "carries an as-of date, a jurisdiction and a requester clearance level" (Section 6.2), so the requester is folded into the context and only clearance is instantiated, although the dimension table lists "clearance, role, purpose." Varying only the requester, with query and context fixed, would show what the third argument adds.

Finding 1 says "no new edge can revoke an inadmissibility verdict." With lineage defined as being "superseded by something in force," an artifact stays inadmissible even after its successor is shown to derive from a retracted source, so a requester can be left with no admissible version at all. Requiring an admissible successor would avoid that but would put negation inside the recursion, and Finding 1 would no longer hold. The authors may intend this conservative outcome, since a superseded rule should not revive when its replacement is defective; if so, the paper should say so. Second, the CQ10b closure asserts ¬(derivedFrom some EvidenceArtifact), which only works for an artifact with no derivation edge at all, the same blunt policy as not_derived in Section 8.3. How is admissibility entailed for a derived artifact whose whole chain is sound? Section 4.1 also names the property supersedes, while Section 4.2 and CQ10b use supersededBy.

From a corrupted-input view, the gate trusts observable metadata, and the only corruption the study applies to it is missingness at a fixed rate (8% per field, 60% for the supersession pointer, Section 6.1). Chosen or forged metadata is different: a publisher who leaves out the derivation pointer, or asserts a global scope or an intact record state, passes not_derived, integrity_intact, and the scope predicates by construction. Section 5.2 notes that asserting admissibility implies complete metadata; it also implies authentic metadata, and the record-state flag cannot show that the text passed to the generator after the gate is the text that was judged. I would like RIER for B, G, and I as curves over the missingness rate, with one condition where the missing pointers fall only on inadmissible artifacts, so that Finding 8 shows

where not_derived and no_retracted_ancestor cross over, if they do. A short threat-model paragraph would also help.

On the statistics, the AUROC near 0.5 on the 800 matched pairs largely follows from their construction (same template, ordinary editorial variation, zero status vocabulary), which is the benign case; under corpus poisoning, inadmissible items are optimized for relevance, and the relation can run the other way. Neither the AUROC values nor the per-dimension range (0.467 to 0.603) has an interval. The parity interval for Admissible Recall@5, [−0.0013, −0.0002], excludes zero and is non-significant only after Holm correction. For RIER, the reference and Snowflake bootstrap p-values (0.734 and 0.256) differ under an identical interval whose lower bound is 0.0000, which suggests the paired differences are exactly zero for most of the 280 confirmatory queries. A count of queries where G and I differ would say more than either p-value. The oracle's Admissible Recall@5 is only 0.118, and its denominator, the number of relevant admissible items per query, is not reported.

Honestly, I would keep the parity experiment and the record of discarded runs as they are; they are what make the narrower claim credible.

Declarations

Potential competing interests: No potential competing interests to declare.