QOM — Quick Ontology Mapping

Marc Ehrig and Steffen Staab

Institute AIFB, University of Karlsruhe

Abstract. (Semi-)automatic mapping — also called (semi-)automatic alignment — of ontologies is a core task to achieve interoperability when two agents or services use different ontologies. In the existing literature, the focus has so far been on improving the quality of mapping results. We here consider QOM, Quick Ontology Mapping, as a way to trade off between effectiveness (i.e. quality) and efficiency of the mapping generation algorithms. We show that QOM has lower run-time complexity than existing prominent approaches. Then, we show in experiments that this theoretical investigation translates into practical benefits. While QOM gives up some of the possibilities for producing high-quality results in favor of efficiency, our experiments show that this loss of quality is marginal.

1 Introduction

Semantic mapping[1] between ontologies is a necessary precondition to establish interoperation between agents or services using different ontologies. In recent years we have seen a range of research work on methods proposing such mappings [1–3]. The focus of the previous work, however, has been laid exclusively on improving the *effectiveness* of the approach (i.e. the quality of proposed mappings such as evaluated against some human judgement given either a posteriori or a priori).

When we tried to apply these methods to some of the real-world scenarios we address in other research contributions [4], we found that existing mapping methods were not suitable for the ontology integration task at hand, as they all neglected *efficiency*. To illustrate our requirements: We have been working in realms where light-weight ontologies are applied such as the ACM Topic hierarchy with its 10^4 concepts or folder structures of individual computers, which corresponded to 10^4 to 10^5 concepts. Finally, we are working with Wordnet exploiting its 10^6 concepts (cf. [5]). When mapping between such light-weight ontologies, the trade-off that one has to face is between effectiveness and efficiency. For instance, consider the knowledge management platform built on a Semantic Web And Peer-to-peer basis in SWAP [4]. It is not sufficient to provide its user with the best possible mapping, it is also necessary to answer his queries within a few seconds — even if two peers use two different ontologies and have never encountered each other before.

In this paper we present an approach that considers both the quality of mapping results as well as the run-time complexity. Our hypothesis is that mapping algorithms may be streamlined such that the loss of quality (compared to a standard baseline) is marginal, but the improvement of efficiency is so tremendous that it allows for the

[1] Frequently also called alignment.

ad-hoc mapping of large-size, light-weight ontologies. To substantiate the hypothesis, we outline a comparison of the worst-case run-time behavior (given in full detail in [6]) and we report on a number of practical experiments. The approaches used for our (unavoidably preliminary) comparison represent different classes of algorithms for ontology mapping. Comparing to these approaches we can observe that our new efficient approach QOM achieves good quality. The complexity of QOM is of $O(n \cdot log(n))$ (measuring with n being the number of the entities in the ontologies) against $O(n^2)$ for approaches that have similar effective outcomes.

The remainder of the paper starts with a clarification of terminology (Section 2). To compare the worst-case run-time behavior of different approaches, we then describe a canonical process for ontology mapping that subsumes the different approaches compared in this paper (Section 3). The process is a core building block for later deriving the run-time complexity of the different mapping algorithms. Section 4 presents our toolbox to analyze these algorithms. In Section 5, different approaches for proposing mappings are described and aligned to the canonical process, one of them being our approach QOM. The way to derive their run-time complexity is outlined in Section 6. Experimental results (Section 7) complement the comparison of run-time complexities. We close this paper with a short section on related work and a conclusion.

2 Terminology

2.1 Ontology

As we currently focus on light-weight ontologies, we build on RDF/S[2] to represent ontologies. To facilitate the further description, we briefly summarize its major primitives and introduce some shorthand notations. An RDF model is described by a set of statements, each consisting of a subject, a predicate and an object. An ontology O is defined by its set of Concepts $\mathcal{C}$ (instances of "rdfs:Class") with a corresponding subsumption hierarchy H_C (a binary relation corresponding to "rdfs:subClassOf"). Relations $\mathcal{R}$ (instances of "rdf:Property") exist between single concepts. Relations are arranged alike in a hierarchy H_R ("rdfs:subPropertyOf"). An entity $i \in \mathcal{I}$ may be an instance of a class $c \in \mathcal{C}$ ("rdf:type"). An instance $i \in \mathcal{I}$ may have one j or many role fillers from $\mathcal{I}$ for a relation r from $\mathcal{R}$. We also call this type of triple (i, r, j) a property instance.

2.2 Mapping

We here define our use of the term "mapping". Given two ontologies O_1 and O_2, mapping one ontology onto another means that for each entity (concept C, relation R, or instance I) in ontology O_1, we try to find a corresponding entity, which has the same intended meaning, in ontology O_2.

Definition 1. *We define an ontology mapping function,* map*, based on the vocabulary,* $\mathcal{E}$*, of all terms* $e \in \mathcal{E}$ *and based on the set of possible ontologies,* $\mathcal{O}$*, as a partial function:*

$$\mathrm{map} : \mathcal{E} \times \mathcal{O} \times \mathcal{O} \rightharpoonup \mathcal{E},$$

[2] http://www.w3.org/RDFS/

with $\forall e \in O_1(\exists f \in O_2 : \text{map}(e, O_1, O_2) = f \vee \text{map}(e, O_1, O_2) = \perp)$.

A term e interpreted in an ontology O is either a concept, a relation or an instance, i.e. $e_{|O} \in \mathcal{C} \cup \mathcal{R} \cup \mathcal{I}$. We usually write e instead of $e_{|O}$ when the ontology O is clear from the context of the writing. We write $\text{map}_{O_1,O_2}(e)$ for $\text{map}(e, O_1, O_2)$. We derive a relation map_{O_1,O_2} by defining $\text{map}_{O_1,O_2}(e, f) \Leftrightarrow \text{map}_{O_1,O_2}(e) = f$. We leave out O_1, O_2 when they are evident from the context and write $\text{map}(e) = f$ and $\text{map}(e, f)$, respectively. Once a (partial) mapping, map, between two ontologies O_1 and O_2 is established, we also say "*entity e is mapped onto entity f*" iff $\text{map}(e, f)$. An entity can either be mapped to at most one other entity. A pair of entities (e, f) that is not yet in map and for which appropriate mapping criteria still need to be tested is called a *candidate mapping*.

2.3 Example

The following example illustrates a mapping. Two ontologies O_1 and O_2 describing the domain of car retailing are given (Figure 1). A reasonable mapping between the two ontologies is given in Table 1 as well as by the dashed lines in the figure.

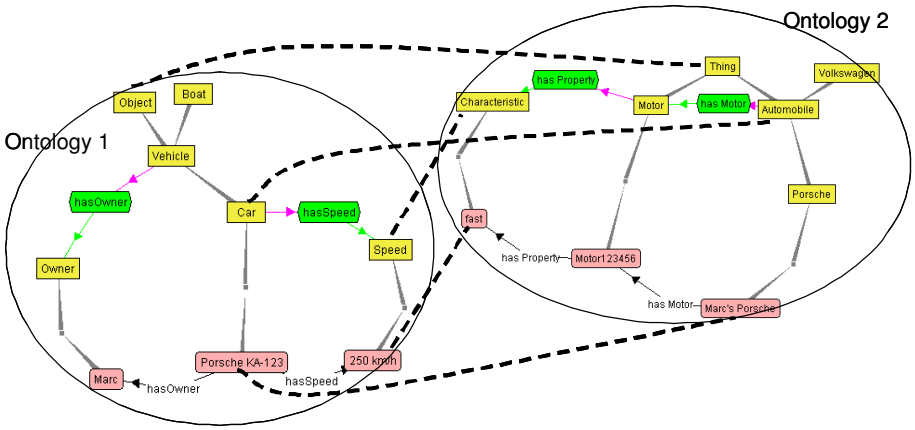

Fig. 1. Example Ontologies and their Mappings

Ontology O_1	Ontology O_2
Object	Thing
Car	Automobile
Porsche KA-123	Marc's Porsche
Speed	Characteristic
250 km/h	fast

Table 1. Mapping Table for Relation $\text{map}_{O_1,O_2}(e, f)$

Apart from one-to-one mappings as investigated in this paper one entity often has to be mapped to a complex composite such as a concatenation of terms (first and last name) or an entity with restrictions (a sports-car is a car going faster than 250 km/h). [7,

8] propose approaches for this. We do not deal with such issues of complete ontology mappings here.

3 Process

We briefly introduce a canonical process that subsumes all the mapping approaches we are aware of.[3] Figure 2 illustrates its six main steps. It is started with two ontologies, which are going to be mapped onto one another, as its input:

1. Feature engineering transforms the initial representation of ontologies into a format digestible for the similarity calculations. For instance, the subsequent mapping process may only work on a subset of RDFS primitives.

2. Selection of Next Search Steps. The derivation of ontology mappings takes place in a search space of candidate mappings. This step may choose, to compute the similarity of a restricted subset of candidate concepts pairs $\{(e, f)|e \in O_1, f \in O_2\}$ and to ignore others.

3. Similarity Computation determines similarity values of candidate mappings.

4. Similarity Aggregation. In general, there may be several similarity values for a candidate pair of entities e, f from two ontologies O_1, O_2, e.g. one for the similarity of their labels and one for the similarity of their relationship to other terms. These different similarity values for one candidate pair must be aggregated into a single aggregated similarity value.

5. Interpretation uses the individual or aggregated similarity values to derive mappings between entities from O_1 and O_2. Some mechanisms here are, to use thresholds for similarity mappings [2], to perform relaxation labelling [3], or to combine structural and similarity criteria.

6. Iteration. Several algorithms perform an iteration over the whole process in order to bootstrap the amount of structural knowledge. Iteration may stop when no new mappings are proposed. Note that in a subsequent iteration one or several of steps 1 through 5 may be skipped, because all features might already be available in the appropriate format or because some similarity computation might only be required in the first round.

Eventually, the output returned is a mapping table representing the relation map_{O_1,O_2}.

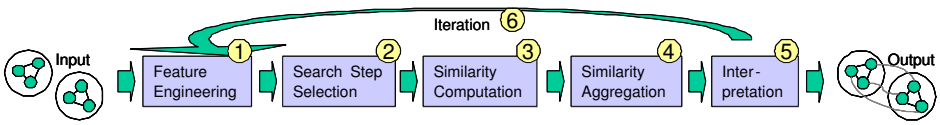

Fig. 2. Mapping Process

[3] The process is inspired by CRISP-DM, http://www.crisp-dm.org/, the CRoss Industry Standard Process for Data Mining.

4 A Toolbox of Data Structures and Methods

The principal idea of this section is to provide a toolbox of data structures and methods common to many approaches that determine mappings. This gives us a least common denominator based on which concrete approaches instantiating the process depicted in Figure 2 can be compared more easily.

4.1 Features of Ontological Entities

To compare two entities from two different ontologies, one considers their characteristics, i.e. their features. The features may be specific for a mapping generation algorithm, in any case the features of ontological entities (of concepts, relations, instances) need to be extracted from extensional and intensional ontology definitions. See also [9] and [10] for an overview of possible features and a classification of them. Possible characteristics include:

- *Identifiers*: i.e. strings with dedicated formats, such as unified resource identifiers (URIs) or RDF labels.
- *RDF/S Primitives*: such as properties or subclass relations
- *Derived Features*: which constrain or extend simple RDFS primitives (e.g. most-specific-class-of-instance)
- *Aggregated Features*: i.e. aggregating more than one simple RDFS primitive, e.g. a sibling is every instance-of the parent-concept of an instance
- *OWL Primitives*: such as an entity being the sameAs another entity
- *Domain Specific Features* are features which only apply to a certain domain with a predefined shared ontology. For instance, in an application where files are represented as instances and the relation hashcode-of-file is defined, we use this feature to compare representations of concrete files.

Example We again refer to the example in Figure 1. The actual feature consists of a juxtaposition of relation name and entity name. The Car concept of ontology 1 is characterized through its (label, Car), the concept which it is linked to through (subclassOf, Vehicle), its (concept sibling, boat), and the (direct property, hasSpeed). Car is also described by its instances through (instance, Porsche KA-123). The relation hasSpeed on the other hand is described through the (domain, Car) and the (range, Speed). An instance would be Porsche KA-123, which is characterized through the instantiated (property instance, (hasOwner, Marc)) and (property instance, (hasSpeed, 250 km/h)).

4.2 Similarity Computation

Definition 2. *We define a similarity measure for comparison of ontology entities as a function as follows (cf. [11]):*

$$\mathrm{sim} : \mathcal{E} \times \mathcal{E} \times \mathcal{O} \times \mathcal{O} \rightarrow [0, 1]$$

Different similarity measures $\mathrm{sim}_k(e, f, O_1, O_2)$ are indexed through a label k. Further, we leave out O_1, O_2 when they are evident from the context and write $\mathrm{sim}_k(e, f)$. The following similarity measures are needed to compare the features of ontological entities at iteration t.

- *Object Equality* is based on existing logical assertions — especially assertions from previous iterations: $\mathrm{sim}_{obj}(a, b) := \{1 \text{ iff } \mathrm{map}_{t-1}(a) = b, 0 \text{ otherwise}\}$
- *Explicit Equality* checks whether a logical assertion already forces two entities to be equal: $\mathrm{sim}_{exp}(a, b) := \{1 \text{ iff } statement(a, \text{“}sameAs'', b), 0 \text{ otherwise}\}$
- *String Similarity* measures the similarity of two strings on a scale from 0 to 1 (cf. [12]) based on Levenshtein's edit distance, ed [13].
  $\mathrm{sim}_{str}(c, d) := max(0, \frac{min(|c|,|d|) - ed(c,d)}{min(|c|,|d|)})$
- *SimSet*: For many features we have to determine to what extent two sets of entities are similar. To remedy the problem, multidimensional scaling [14] measures how far two entities are from all other entities and assumes that if they have very similar distances to all other entities, they must be very similar:
  $\mathrm{sim}_{set}(E, F) = \frac{\sum_{e \in E} \boldsymbol{e}}{|E|} \cdot \frac{\sum_{f \in F} \boldsymbol{f}}{|F|}$
  with $\boldsymbol{e} = (\mathrm{sim}(e, e_1), \mathrm{sim}(e, e_2), \ldots, \mathrm{sim}(e, f_1), \mathrm{sim}(e, f_2), \ldots)$, $\boldsymbol{f}$ analogously.

These measures are all input to the similarity aggregation.

4.3 Similarity Aggregation

Similarities are aggregated by:

$$\mathrm{sim}_{agg}(e, f) = \frac{\sum_{k=1\ldots n} w_k \cdot \mathrm{adj}(\mathrm{sim}_k(e, f))}{\sum_{k=1\ldots n} w_k}$$

with w_k being the weight for each individual similarity measure, and adj being a function to transform the original similarity value ($adj : [0, 1] \rightarrow [0, 1]$), which yields better results.

4.4 Interpretation

From the similarity values we derive the actual mappings. The basic idea is that each entity may only participate in one mapping and that we assign mappings based on a threshold t and a greedy strategy that starts with the largest similarity values first. Ties are broken arbitrarily by $\tilde{argmax}_{(g,h)}$, but with a deterministic strategy.

$P(\bot, \bot, E \cup \{\bot\}, E \cup \{\bot\})$.
$P(g, h, U \backslash \{e\}, V \backslash \{f\}) \leftarrow P(e, f, U, V) \wedge \mathrm{sim}(g, h) > t$
$\quad \wedge (g, h) = \tilde{\mathrm{argmax}}_{(g,h) \in U \backslash \{e\} \times V \backslash \{f\}} \mathrm{sim}_{\mathrm{agg}}(g, h)$.
$\mathrm{map}(e, f) \leftarrow \exists X_1, X_2 P(e, f, X_1, X_2) \wedge (e, f) \neq (\bot, \bot)$.

5 Approaches to Determine Mappings

In the following we now use the toolbox, and extend it, too, in order to define a range of different mapping generation approaches. In the course of this section we present our novel Quick Ontology Mapping approach — QOM.

5.1 Standard Mapping Approaches

Our Naive Ontology Mapping (NOM)[9] constitutes a straight forward baseline for later comparisons. It is defined by the steps of the process model as follows. Where appropriate we point to related mapping approaches and briefly describe the difference in comparison to NOM.

1. Feature Engineering Firstly, the ontologies have to be represented in RDFS. We use features as shown in Section 4.1.
For PROMPT and Anchor-PROMPT [2] any ontology format is suitable as long as it can be used in the Protege environment. GLUE [3] learns in advance, based on a sample mapping set, a similarity estimator to identify equal instances and concepts.

2. Search Step Selection All entities of the first ontology are compared with all entities of the second ontology. Any pair is treated as a candidate mapping.
This is generally the same for other mapping approaches, though the PROMPT algorithm can be implemented more efficiently by sorting the labels first, thus only requiring the comparison of two neighboring elements in the list.

3. Similarity Computation The similarity computation between an entity of O_1 and an entity of O_2 is done by using a wide range of similarity functions. Each similarity function is based on a feature (Section 4.1) of both ontologies and a respective similarity measure (Section 4.2). For NOM they are shown in Table 2.
The PROMPT system determines the similarity based on the exact equality (not only similarity) of labels. Anchor-PROMPT adds structural components. In GLUE the similarity is gained using the previously learned Similarity Estimator. It further adds other features using Relaxation Labelling based on the intuition that mappings of a node are typically influenced by the node's neighborhood.

4. Similarity Aggregation NOM emphasizes high individual similarities and deemphasizes low individual similarities by weighting individual similarity results with a sigmoid function first and summing the modified values then. To produce an aggregated similarity (cf. Section 4.2) NOM applies $adj(x) = \frac{1}{1+e^{-5(x-0.5)}}$. Weights w_k are assigned by manually maximizing the f-measure on overall training data from different test ontologies.
In systems with one similarity value such as PROMPT or GLUE this step does not apply.

5. Interpretation NOM interpretes similarity results by two means. First, it applies a threshold to discard spurious evidence of similarity. Second, NOM enforces bijectivity of the mapping by ignoring candidate mappings that would violate this constraint and by favoring candidate mappings with highest aggregated similarity scores.
As PROMPT is semi-automatic this step is less crucial. It presents all pairs with a similarity value above a relatively low threshold value and the users can decide to carry out the merging step or not. The relaxation labelling process of GLUE can also be seen as a kind of interpretation.

6. Iteration The first round uses only the basic comparison method based on labels and string similarity to compute the similarity between entities. By doing the computation in several rounds one can access the already computed pairs and use more sophisticated structural similarity measures. Therefore, in the second round and thereafter NOM re-

Comparing	No.	Feature	Similarity Measure
Concepts	1	(label,X_1)	string similarity(X_1, X_2)
	2	(URI_1)	string equality(URI_1, URI_2)
	3	(X_1,sameAs,X_2) relation	explicit equality(X_1, X_2)
	4	(direct properties,Y_1)	SimSet(Y_1, Y_2)
	5	all (inherited properties,Y_1)	SimSet(Y_1, Y_2)
	6	all (super-concepts,Y_1)	SimSet(Y_1, Y_2)
	7	all (sub-concepts,Y_1)	SimSet(Y_1, Y_2)
	8	(concept siblings,Y_1)	SimSet(Y_1, Y_2)
	9	(direct instances,Y_1)	SimSet(Y_1, Y_2)
	10	(instances,Y_1)	SimSet(Y_1, Y_2)
Relations	1	(label,X_1)	string similarity(X_1, X_2)
	2	(URI_1)	string equality(URI_1, URI_2)
	3	(X_1,sameAs,X_2) relation	explicit equality(X_1, X_2)
	4	(domain,X_d1) and (range,X_r1)	object equality(X_{d1}, X_{d2}), (X_{r1}, X_{r2})
	5	all (super-properties,Y_1)	SimSet(Y_1, Y_2)
	6	all (sub-properties,Y_1)	SimSet(Y_1, Y_2)
	7	(property siblings,Y_1)	SimSet(Y_1, Y_2)
	8	(property instances,Y_1)	SimSet(Y_1, Y_2)
Instances	1	(label,X_1)	string similarity(X_1, X_2)
	2	(URI_1)	string equality(URI_1, URI_2)
	3	(X_1,sameAs,X_2) relation	explicit equality(X_1, X_2)
	4	all (parent-concepts,Y_1)	SimSet(Y_1, Y_2)
	5	(property instances,Y_1)	SimSet(Y_1, Y_2)
Property-Instances	1	(domain,X_d1) and (range,X_r1)	object equality(X_{d1}, X_{d2}), (X_{r1}, X_{r2})
	2	(parent property,Y_1)	SimSet(Y_1, Y_2)

Table 2. Features and Similarity Measures for Different Entity Types Contributing to Aggregated Similarity in NOM. The corresponding ontology is indicated through an index.

lies on all the similarity functions listed in Table 2.
PROMPT also requires these iterations after feedback has been given by the user. The GLUE system heavily relies on iterations for the relaxation labelling process. Both do not change the strategies during the iterations.

5.2 QOM — Quick Ontology Mapping

The goal of this paper is to present an efficient mapping algorithm. For this purpose, we optimize the effective, but inefficient NOM approach towards our goal. The outcome is QOM — Quick Ontology Mapping. We would also like to point out that the efficiency gaining steps can be applied to other mapping approaches as well.

1. Feature Engineering Like NOM, QOM exploits RDF triples.

2. Search Step Selection A major ingredient of run-time complexity is the number of candidate mapping pairs which have to be compared to actually find the best mappings. Therefore, we use heuristics to lower the number of candidate mappings. Fortunately we can make use of ontological structures to classify the candidate mappings into promising and less promising pairs.

In particular we use a dynamic programming approach [15]. In this approach we have two main data structures. First, we have candidate mappings which ought to be investigated. Second, an agenda orders the candidate mappings, discarding some of them entirely to gain efficiency. After the completion of the similarity analysis and their interpretation new decisions have to be taken. The system has to determine which candidate mappings to add to the agenda for the next iteration. The behavior of initiative and ordering constitutes a search strategy.

We suggest the subsequent strategies to propose new candidate mappings for inspection:

Random A simple approach is to limit the number of candidate mappings by selecting either a fixed number or percentage from all possible mappings.

Label This restricts candidate mappings to entity pairs whose labels are near to each other in a sorted list. Every entity is compared to its "label"-neighbors.

Change Propagation QOM further compares only entities for which adjacent entities were assigned new mappings in a previous iteration. This is motivated by the fact that every time a new mapping has been found, we can expect to also find similar entities adjacent to these found mappings. Further, to prevent very large numbers of comparisons, the number of pairs is restricted.

Hierarchy We start comparisons at a high level of the concept and property taxonomy. Only the top level entities are compared in the beginning. We then subsequently descend the taxonomy.

Combination The combined approach used in QOM follows different optimization strategies: it uses a label subagenda, a randomness subagenda, and a mapping change propagation subagenda. In the first iteration the label subagenda is pursued. Afterwards we focus on mapping change propagation. Finally we shift to the randomness subagenda, if the other strategies do not identify sufficiently many correct mapping candidates.

With these multiple agenda strategies we only have to check a fixed and restricted number of mapping candidates for each original entity.[4] Please note that the creation of the presented agendas does require processing resources itself.

3. Similarity Computation QOM, just like NOM, is based on a wide range of ontology feature and heuristic combinations. In order to optimize QOM, we have restricted the range of costly features as specified in Table 3. In particular, QOM avoids the complete pair-wise comparison of trees in favor of a(n incomplete) top-down strategy. The marked comparisons in the table were changed from features which point to complete inferred sets to features only retrieving limited size direct sets.

4. Similarity Aggregation The aggregation of single methods is only performed once per candidate mapping and is therefore not critical for the overall efficiency. Therefore, QOM works like NOM in this step.

5. Interpretation Also the interpretation step of QOM is the same as in NOM.

6. Iteration QOM iterates to find mappings based on lexical knowledge first and based on knowledge structures later.

[4] We have also explored a number of other strategies or combinations of strategies with simple data sets but they did not outperform results of QOM presented here.

Comparing	Change	Feature	Similarity Measure
Concepts	5	all (inherited properties, Y_1)	SimSet(Y_1, Y_2)
	$\longrightarrow$ 5a	(properties of direct super-concepts, Y_1)	SimSet(Y_1, Y_2)
	6	all (inherited super-concepts, Y_1)	SimSet(Y_1, Y_2)
	$\longrightarrow$ 6a	(direct super-concepts, Y_1)	SimSet(Y_1, Y_2)
	7	all (inherited sub-concepts, Y_1)	SimSet(Y_1, Y_2)
	$\longrightarrow$ 7a	(direct sub-concepts, Y_1)	SimSet(Y_1, Y_2)
	10	all (inherited instances, Y_1)	SimSet(Y_1, Y_2)
	$\longrightarrow$ 10a	(instances of direct sub-concepts, Y_1)	SimSet(Y_1, Y_2)
Relations	5	all (inherited super-properties, Y_1)	SimSet(Y_1, Y_2)
	$\longrightarrow$ 5a	(direct super-properties, Y_1)	SimSet(Y_1, Y_2)
	6	all (inherited sub-properties, Y_1)	SimSet(Y_1, Y_2)
	$\longrightarrow$ 6a	(direct sub-properties, Y_1)	SimSet(Y_1, Y_2)
Instances	4	all (inherited parent-concepts, Y_1)	SimSet(Y_1, Y_2)
	$\longrightarrow$ 4a	(direct parent-concepts, Y_1)	SimSet(Y_1, Y_2)

Table 3. Features and Similarity Measures for Different Entity Types Contributing to Aggregated Similarity in QOM. The lower case "a" indicates that the feature has been modified for efficiency considerations.

In all our tests we have found that after ten rounds hardly any further changes occur in the mapping table. This is independent from the actual size of the involved ontologies. QOM therefore restricts the number of runs.

Assuming that ontologies have a fixed percentage of entities with similar lexical labels, we will easily find their correct mappings in the first iteration. These are further evenly distributed over the two ontologies, i.e. the distance to the furthest not directly found mapping is constant. Through the change propagation agenda we pass on to the next adjacent mapping candidates with every iteration step. The number of required iterations remains constant; it is independent from the size of the ontologies.

6 Comparing Run-time Complexity

We determine the worst-case run-time complexity of the algorithms to propose mappings as a function of the size of the two given ontologies. Thereby, we wanted to base our analysis on realistic ontologies and not on artifacts. We wanted to avoid the consideration of large ontologies with n leaf concepts but a depth of the concept hierarchy H_C of $n - 1$. [16] have examined the structure of a large number of ontologies and found, that concept hierarchies on average have a branching factor of around 2 and that the concept hierarchies are neither extremely shallow nor extremely deep. The actual branching factor can be described by a power law distribution. Hence, in the following we base our results on their findings.

Theorem 1. *The worst case run-time behaviors of NOM, PROMPT, Anchor-PROMPT, GLUE and QOM are given by the following table:*

NOM	$O(n^2 \cdot log^2(n))$
PROMPT[5]	$O(n \cdot log(n))$
Anchor-PROMPT	$O(n^2 \cdot log^2(n))$
GLUE[6]	$O(n^2)$
QOM	$O(n \cdot log(n))$

Proof Sketch 1 *The different algorithmic steps contributing to complexity*[7] *are aligned to the canonical process of Section 3.*

*For each of the algorithms, one may then determine the costs of each step. First, one determines the cost for feature engineering (*feat*). The second step is the search step i.e. candidate mappings selection (*sele*). For each of the selected candidate mappings (*comp*) we need to compute* k *different similarity functions* sim_k *and aggregate them (*agg*). The number of entities involved and the complexity of the respective similarity measure affect the run-time performance. Subsequently the interpretation of the similarity values with respect to mapping requires a run-time complexity of* inter. *Finally we have to iterate over the previous steps multiple times (*iter*).*

Then, the worst case run-time complexity is defined for all approaches by:

$$c = (feat + sele + comp \cdot (\textstyle\sum_k sim_k + agg) + inter) \cdot iter$$

Depending on the concrete values that show up in the individual process steps the different run-time complexities are derived in detail in [6].

7 Empirical Evaluation and Results

In this section we show that the worst case considerations carry over to practical experiments and that the quality of QOM is only negligibly lower than the one of other approaches. The implementation itself was coded in Java using the KAON-framework[8] for ontology operations.

7.1 Test Scenario

Metrics We use standard information retrieval metrics to assess the different approaches (cf. [17]):

Precision $p = \frac{\#correct\_found\_mapping}{\#found\_mappings}$
Recall $r = \frac{\#correct\_found\_mappings}{\#existing\_mappings}$
F-Measure $f_1 = \frac{2pr}{p+r}$

[5] This complexity assumes an ideal implementation of PROMPT using a sorted list. The tool itself requires $O(n^2)$.

[6] This result is based on optimistic assumptions about the learner.

[7] In this paper we assume that the retrieval of a statement of an ontology entity from a database can be done in constant access time, independent of the ontology size, e.g. based on sufficient memory and a hash function.

[8] http://kaon.semanticweb.org/

Data Sets Three separate data sets were used for evaluation purposes. As real world ontologies and especially their mappings are scarce, students were asked to independently create and map ontologies.[9]

Russia 1 In this first set we have two ontologies describing Russia. The students created the ontologies with the objectives to represent the content of two independent travel websites about Russia. These ontologies have approximately 400 entities each, including concepts, relations, and instances. The total number of possible mappings is 160, which the students have assigned manually.

Russia 2 The second set again covers Russia, but the two ontologies are more difficult to map. After their creation they have been altered by deleting entities and changing the labels at random. They differ substantially in both labels and structure. Each ontology has 300 entities with 215 possible mappings, which were captured during generation.

Tourism Finally, the participants of a seminar created two ontologies which separately describe the tourism domain of Mecklenburg-Vorpommern. Both ontologies have an extent of about 500 entities. No instances were modelled with this ontology though, they only consist of concepts and relations. The 300 mappings were created manually.

Strategies We evaluated the mapping strategies described in the previous sections:

- PROMPT — As the PROMPT algorithm is rather simple and fast we use it as a baseline to evaluate the speed. The empirical evaluation is based on the actual implementation of PROMPT rather than its theoretic potential, as described in the previous section.
- NOM / Anchor-PROMPT — Naive Ontology Mapping is an approach making use of a wide range of features and measures. Therefore it reaches high levels of effectiveness and represents our quality baseline. In terms of structural information used and complexity incurred it is similar to Anchor-PROMPT.
- QOM — Quick Ontology Mapping is our novel approach focusing on efficiency.

To circumvent the problem of having semi-automatic merging tools (PROMPT and Anchor-PROMPT) in our fully automatic mapping tests, we assumed that every proposition of the system is meaningful and correct. Further, as we had difficulties in running Anchor-PROMPT with the size of the given data sets, we refer to the results of the somewhat similar NOM. For GLUE we face another general problem. The algorithm has a strong focus on example instance mappings. As we can not provide this, we refrained from running the tests on a poorly trained estimator which would immediately result in poor quality results.

7.2 Results and Discussion

We present the results of the strategies on each of the data sets in Figures 3 and 4. The tourism dataset shows similar characteristics as Russia 1 and is therefore not plotted. The x-axis shows the elapsed time on a logarithmic scale, the y-axis corresponds to the f-measure. The symbols represent the result after each iteration step.

[9] The datasets are available from http://www.aifb.uni-karlsruhe.de/WBS/meh/mapping/.

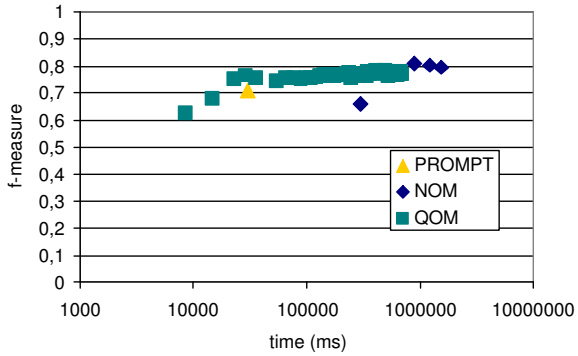

Fig. 3. Mapping quality reached over time with Russia 1 ontologies.

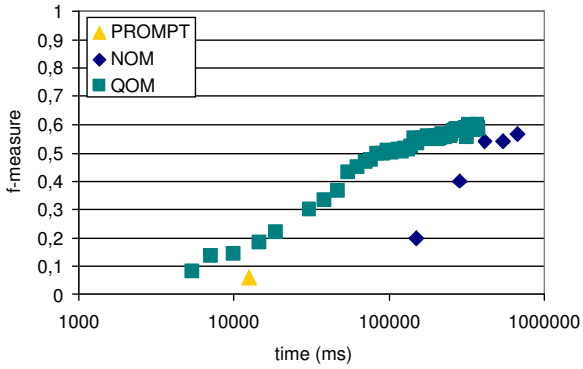

Fig. 4. Mapping quality reached over time with Russia 2 ontologies.

Depending on the scenario PROMPT reaches good results within a short period of time. Please notice that for ontologies with a small number of similar labels (Figure 4) this strategy is not satisfactory (f-measure 0.06). In contrast, the f-measure value of the NOM strategy rises slowly but reaches high absolute values of up to 0.8. Unfortunately it requires a lot of time. Finally the QOM Strategy is plotted. It reaches high quality levels very quickly. In terms of absolute values it also seems to reach the best quality results of all strategies. This appears to be an effect of QOM achieving an about 20 times higher number of iterations than NOM within the given time frame.

Lessons Learned. We had the hypothesis that faster mapping results can be obtained with only a negligible loss of quality. We here briefly present the bottom line of our considerations in this paper:

1. Optimizing the mapping approach for efficiency — like QOM does — decreases the overall mapping quality. If ontologies are not too large one might prefer to rather avoid this.
2. Labels are very important for mapping, if not the most important feature of all, and alone already return very satisfying results.

3. Using an approach combining many features to determine mappings clearly leads to significantly higher quality mappings.
4. The Quick Ontology Mapping approach shows very good results. Quality is lowered only marginally, thus supporting our hypothesis.
5. QOM is faster than standard prominent approaches by a factor of 10 to 100 times.

Recapitulating we can say that our mapping approach is very effective and efficient.

8 Related Work

We only present closely related work not yet mentioned in this paper.

Various authors have tried to find a general description of similarity with several of them being based on knowledge networks. [18] give a general overview of similarity.

Original work on mapping is presented by [19] in their tool ONION, which uses inferencing to execute mappings, but is based on manually assigned mappings or very simple heuristics. An interesting approach for schema and ontology mapping is presented by [20]. Explicit semantic rules are added for consideration. A SAT solver is used to prevent mappings to imply semantical contradictions.

Despite the large number of related work on effective mapping already mentioned throughout this paper, there are very few approaches raising the issue of efficiency.

Apart from the ontology domain research on mapping and integration has been done in various computer science fields. [1] present an approach to integrate documents from different sources into a master catalog. There has also been research on efficient schema and instance integration within the database community. [21] is a good source for an overview. Due to the different domain comparisons with our approach are very difficult.

9 Conclusion

The problem of mapping two ontologies effectively and efficiently arises in many application scenarios [4, 5]. We have devised a generic process model to investigate and compare different approaches that generate ontology mappings. In particular, we have developed an original method, QOM, for identifying mappings between two ontologies. We have shown that it is on a par with other good state-of-the-art algorithms concerning the quality of proposed mappings, while outperforming them with respect to efficiency — in terms of run-time complexity ($O(n \cdot log(n))$ instead of $O(n^2)$) and in terms of the experiments we have performed (by a factor of 10 to 100).

Acknowledgements Research reported in this paper has been partially financed by the EU in the IST projects WonderWeb (IST-2001-33052), SWAP (IST-2001-34103) and SEKT (IST-2003-506826).

References

1. Agrawal, R., Srikant, R.: On integrating catalogs. In: Proceedings of the tenth international conference on World Wide Web, ACM Press (2001) 603–612

2. Noy, N.F., Musen, M.A.: The PROMPT suite: interactive tools for ontology merging and mapping. International Journal of Human-Computer Studies **59** (2003) 983–1024
3. Doan, A., Domingos, P., Halevy, A.: Learning to match the schemas of data sources: A multistrategy approach. VLDB Journal **50** (2003) 279–301
4. Ehrig, M., Haase, P., van Harmelen, F., Siebes, R., Staab, S., Stuckenschmidt, H., Studer, R., Tempich, C.: The SWAP data and metadata model for semantics-based peer-to-peer systems. In: Proceedings of MATES-2003. First German Conference on Multiagent Technologies. LNAI, Erfurt, Germany, Springer (2003)
5. Hotho, A., Staab, S., Stumme, G.: Ontologies improve text document clustering. In: Proceedings of the International Conference on Data Mining — ICDM-2003, IEEE Press (2003)
6. Ehrig, M., Staab, S.: Quick ontology mapping with QOM. Technical report, University of Karlsruhe, Institute AIFB (2004) http://www.aifb.uni-karlsruhe.de/WBS/meh/mapping/.
7. Do, H., Rahm, E.: COMA - a system for flexible combination of schema matching approaches. In: Proceedings of the 28th VLDB Conference, Hong Kong, China (2002)
8. Dhamankar, R., Lee, Y., Doan, A., Halevy, A., Domingos, P.: imap: discovering complex semantic matches between database schemas. In: Proceedings of the 2004 ACM SIGMOD international conference on Management of data. (2004) 383–394
9. Ehrig, M., Sure, Y.: Ontology mapping - an integrated approach. In Bussler, C., Davis, J., Fensel, D., Studer, R., eds.: Proceedings of the 1st ESWS. Volume 3053 of Lecture Notes in Computer Science., Heraklion, Greece, Springer Verlag (2004) 76–91
10. Euzenat, J., Valtchev, P.: An integrative proximity measure for ontology alignment. In Doan, A., Halevy, A., Noy, N., eds.: Proceedings of the Semantic Integration Workshop at ISWC-03. (2003)
11. Bisson, G.: Why and how to define a similarity measure for object based representation systems. Towards Very Large Knowledge Bases (1995) 236–246
12. Maedche, A., Staab, S.: Measuring similarity between ontologies. In: Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW), Springer (2002)
13. Levenshtein, I.V.: Binary codes capable of correcting deletions, insertions, and reversals. Cybernetics and Control Theory (1966)
14. Cox, T., Cox, M.: Multidimensional Scaling. Chapman and Hall (1994)
15. Boddy, M.: Anytime problem solving using dynamic programming. In: Proceedings of the Ninth National Conference on Artificial Intelligence, Anaheim, California, Shaker Verlag (1991) 738–743
16. Tempich, C., Volz, R.: Towards a benchmark for semantic web reasoners - an analysis of the DAML ontology library. In Sure, Y., ed.: Evaluation of Ontology-based Tools (EON2003) at Second International Semantic Web Conference (ISWC 2003). (2003)
17. Do, H., Melnik, S., Rahm, E.: Comparison of schema matching evaluations. In: Proceedings of the second int. workshop on Web Databases (German Informatics Society). (2002)
18. Rodrguez, M.A., Egenhofer, M.J.: Determining semantic similarity among entity classes from different ontologies. IEEE Transactions on Knowledge and Data Engineering (2000)
19. Mitra, P., Wiederhold, G., Kersten, M.: A graph-oriented model for articulation of ontology interdependencies. Lecture Notes in Computer Science **1777** (2000) 86+
20. Bouquet, P., Magnini, B., Serafini, L., Zanobini, S.: A SAT-based algorithm for context matching. In: IV International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT'2003), Stanford University (CA, USA) (2003)
21. McCallum, A., Nigam, K., Ungar, L.H.: Efficient clustering of high-dimensional data sets with application to reference matching. In: Knowledge Discovery and Data Mining. (2000) 169–178