MapPSO Results for OAEI 2009

Jürgen Bock[1], Peng Liu[1], and Jan Hettenhausen[2]

[1] FZI Forschungszentrum Informatik an der Universität Karlsruhe, Germany
{bock, pengliu}@fzi.de
[2] Griffith University, Institute for Integrated and Intelligent Systems, Brisbane, Australia
j.hettenhausen@griffith.edu.au

Abstract. This paper presents and discusses the results of the latest developments of the MapPSO system, which is an ontology alignment approach that is based on discrete particle swarm optimisation. Firstly it is recalled, how the algorithm approaches the ontology matching task as an optimisation problem, and how the specific technique of particle swarm optimisation is applied. Secondly, the results are discussed, which were achieved for the Benchmark data set of the 2009 Ontology Alignment Evaluation Initiative.

1 Presentation of the system

With last year's OAEI campaign the MapPSO system (Ontology **Map**ping by **P**article **S**warm **O**ptimisation) has been introduced [1] as a novel research prototype, which is expected to become a highly scalable, massively parallel tool for ontology alignment. In the following subsection the basic idea of this approach will be sketched.

1.1 State, purpose, general statement

The MapPSO algorithm is being developed for the purpose of aligning large ontologies. It is motivated by the observation that ontologies and schema information such as thesauri or dictionaries are not only getting numerous on the web, but also are becoming increasingly large in terms of the number of classes/concepts and properties/relations. This development raises the need for highly scalable tools to provide interoperability and integration of various heterogeneous sources. On the other hand the emergence of parallel architectures provide the basis for highly parallel and thus scalable algorithms which need to be adapted to these architectures.

The presented MapPSO method regards the ontology alignment problem as an optimisation problem which allows for the adaptation of a discrete variant of particle swarm optimisation [2, 3], a population based optimisation paradigm inspired by social interaction between swarming animals. Particularly the population based structure of this method provides high scalability on parallel systems. Particle swarm optimisation furthermore belongs to the group of anytime algorithms, which allow for interruption at any time and will provide the best answer being available at that time. Particularly this property might be interesting when an alignment problem is subject to certain time constraints.

Compared to the first version of the system that participated in last year's OAEI campaign, some adaptation have been made with particular respect to the base matchers used. More precisely, the existing base matchers have been improved, and new base matchers have been applied, in order to improve the quality of the alignments discovered by MapPSO. Section 2 shows the improvements compared to OAEI 2008.

1.2 Specific techniques used

MapPSO utilises a discrete particle swarm optimisation (DPSO) algorithm, based in parts on the DPSO developed by Correa *et al.* [2, 3], to tackle the ontology matching problem as an optimisation problem. The core element of this optimisation problem is the objective function which supplies a fitness value for each candidate alignment. To find solutions for the optimisation problem, MapPSO simulates a set of particles whereby each particle is a candidate alignment comprising a set of initially random mappings. (Currently only 1:1 alignments are supported.) Each of these particles maintains a memory of previously found good mappings (*personal best*) and the swarm maintains a collective memory of the best known alignment so far (*global best*). In each iteration, particles are updated by changing their sets of correspondences in a guided random manner. Correspondences which are also present in the global best set and personal best set are more likely to be kept, as are those with a very good evaluation. Worst Correspondences are more likely to be removed and replaced with other correspondences which are random recommended from best alignment(*personal best* and *global best*) and random created according to left available entities. Each candidate alignment of two ontologies is scored based on the sum of quality measures of the single correspondences. The currently best alignment is the one with the best known fitness rating according to these criteria. According to this revisit of the ontology matching problem, a particle swarm can be applied to search for the optimal alignment.

For each correspondence the quality score is calculated based on an aggregation of scores from a configurable set of base matchers. Each base matcher provides a distance measure for each correspondence. Currently the following base matchers are used:

- SMOA string distance [4] for entity names
- SMOA string distance for entity labels
- WordNet distance for entity names
- WordNet distance for entity labels
- Vector space similarity [5] for entity comments
- Hierarchy distance to propagate similarity of super/subclasses and super/subproperties
- Structural similarity of classes derived from properties that have them as domain or range classes
- Structural similarity of properties derived from their domain and range classes
- Similarity of classes derived from individuals that are instances of them
- Similarity of properties derived from individuals that are subjects or objects of them
- Similarity of individuals derived from property assertions, in particular the following:
  - values of data properties, the resp. individual is asserted to

- object (individuals) of object properties, the resp. individual is asserted to as subject
- subject (individuals) of object properties, the resp. individual is asserted to as object

For each correspondence the available base distances are aggregated by applying a weighted average operator. Hereby a fixed weight is assigned to each base distance. However, the weight configuration is automatically adjusted before the alignment process, according to the ontology characteristics. By this analysis those characteristics are determined that are most promising for detecting similarities. The evaluation of the overall alignment of each particle is computed by aggregating all its correspondence distances. In the current implementation each particle runs in a separate thread and all fitness calculations and particle updates are performed in parallel. The only sequential portion on the algorithm is the synchronisation after each iteration to acquire the fitness value from each particle and determine the currently global best alignment.

1.3 Adaptations made for the evaluation

Since MapPSO is an early prototype, the OAEI Benchmark test data is used during the development process. No specific adaptations have been made.

1.4 Link to the system and parameters file

The release of MapPSO (`MapPSO.jar`) and the parameter file `params.xml` used for OAEI 2009 are located at `https://sourceforge.net/projects/mappso/files/` in the folder `oaei2009`.

1.5 Link to the set of provided alignments (in align format)

The alignments of the OAEI 2009 benchmark data set as provided by MapPSO are located in the file `alignments.zip` at `https://sourceforge.net/projects/mappso/files/`.

2 Results

The MapPSO system participated only in the benchmarks track this year.

The algorithm is highly adjustable via its parameter file and can be tuned to perform well on specific problems, as well as to perform well for precision or recall. To obtain the results presented in Tab. 1 a compromised parameter configuration was used.

2.1 benchmark

The Benchmark test case is designed to provide a number of data sets systematically revealing strengths and weaknesses of the matching algorithm. In the case of MapPSO the experiences were as follows.

Note, that in the results where computed without consulting WordNet in order to improve run-time performance.

For tests **101–104** MapPSO achieves precision and recall values of 100 %. Since the ontologies in those tests have complete information, which can used for alignment. The results have slightly improved compared to the results from 2008.

As for tests **201–210** results are slightly worse than for tests 101–104, since by each test, one or more types of linguistic information are lost, so the system has to rely on other information and on different base matchers resp. in order to determine the similarity of entities. The quality of the alignment decreases with the number of features that provide linguistic features to exploit. In particular for test 202, all names, labels and comments are unavailable, the system achieves about 63 % precision and recall by using solely structural and semantic information. However, with newly added base matchers which respect ABox information in ontologies, the results for tests 201-210 are much improved as last year.

In tests **221–247**, where the structure of the ontologies varies, the results are similar to the 10x tests. Since the linguistic features can be used by MapPSO, which is still the main focus of the current implementation of MapPSO.

The tests **248–266** combine linguistic and structural problems. As the results show, the quality of the alignments is decreasing with the decreasing number of features available in the ontologies. The results of some tests are slightly worse as 2008, for instance 249-2. The reason is possibly the using of weighted average operator instead of ordered weighted average operator and deactivating WordNet distance.

For the real-world tests **301–304**, no uniform results can be derived as the algorithm's precision and recall values vary between 0 and 60 %.

All together, results of our system MapPSO in 2009 is significantly improved compared to the previous version in 2008, but since the test is run without WordNet there are some tests with worse results.

3 General comments

In the following we will provide a few statements on our experiences from participating in the OAEI 2008 competition and briefly discuss future work on the MapPSO algorithm.

3.1 Comments on the results

Firstly it shall be noted that MapPSO is a non-deterministic method and therefore on a set of independent runs the quality of the results and the number of mappings in the alignments will be subject to slight fluctuations.

3.2 Discussions on the way to improve the proposed system

With the latest version of MapPSO several new base matchers have been applied in the system, which significantly improved the quality of the results. In particular, the system makes use of *lexical*, *linguistic*, *structural*, and to a certain extent *semantic*

Table 1. Results of MapPSO in the OAEI 2009 benchmark data set.

Test Name	Precision	Recall	Test Name	Precision	Recall	Test Name	Precision	Recall
101	1	1	246	0.97	1	257	0.24	0.24
103	1	1	247	0.85	0.88	257-2	0.88	0.88
104	1	1	248	0.61	0.61	257-4	0.94	0.94
201	1	1	248-2	0.61	0.61	257-6	0.61	0.61
201-2	1	1	248-4	0.58	0.58	257-8	0.52	0.52
201-4	0.98	0.98	248-6	0.58	0.58	258	0.1	0.1
201-6	1	1	248-8	0.58	0.58	258-2	0.28	0.28
201-8	1	1	249	0.04	0.04	258-4	0.17	0.17
202	0.64	0.64	249-2	0.3	0.3	258-6	0.07	0.08
202-2	0.94	0.94	249-4	0.23	0.23	258-8	0.12	0.12
202-4	0.7	0.7	249-6	0.12	0.12	259	0.04	0.04
202-6	0.86	0.86	249-8	0.1	0.1	259-2	0.23	0.23
202-8	0.69	0.69	250	0.39	0.39	259-4	0.22	0.22
203	1	1	250-2	1	1	259-6	0.23	0.23
204	1	1	250-4	0.79	0.79	259-8	0.21	0.21
205	1	0.99	250-6	0.55	0.55	260	0.13	0.14
206	1	0.99	250-8	0.48	0.48	260-2	0.77	0.79
207	1	0.99	251	0.58	0.58	260-4	0.6	0.62
208	0.97	0.97	251-2	0.87	0.87	260-6	0.37	0.38
209	0.68	0.67	251-4	0.72	0.72	260-8	0.33	0.34
210	0.7	0.7	251-6	0.58	0.58	261	0.12	0.12
221	1	1	251-8	0.57	0.57	261-2	0.47	0.48
222	1	1	252	0.45	0.45	261-4	0.59	0.61
223	0.97	0.97	252-2	0.77	0.77	261-6	0.53	0.55
224	1	1	252-4	0.76	0.76	261-8	0.5	0.52
225	1	1	252-6	0.77	0.77	262	0.06	0.06
228	1	1	252-8	0.84	0.84	262-2	0.76	0.76
230	0.91	0.93	253	0.06	0.06	262-4	0.58	0.58
231	1	1	253-2	0.18	0.18	262-6	0.45	0.45
232	1	1	253-4	0.06	0.06	262-8	0.27	0.27
233	1	1	253-6	0.03	0.03	265	0.1	0.1
236	1	1	253-8	0.09	0.09	266	0.06	0.06
237	0.99	1	254	0.18	0.18	301	0.47	0.44
238	0.96	0.96	254-2	0.7	0.7	302	NaN	0
239	0.97	1	254-4	0.48	0.48	303	NaN	0
240	0.82	0.85	254-6	0.3	0.3	304	0.59	0.53
241	1	1	254-8	0.12	0.12			

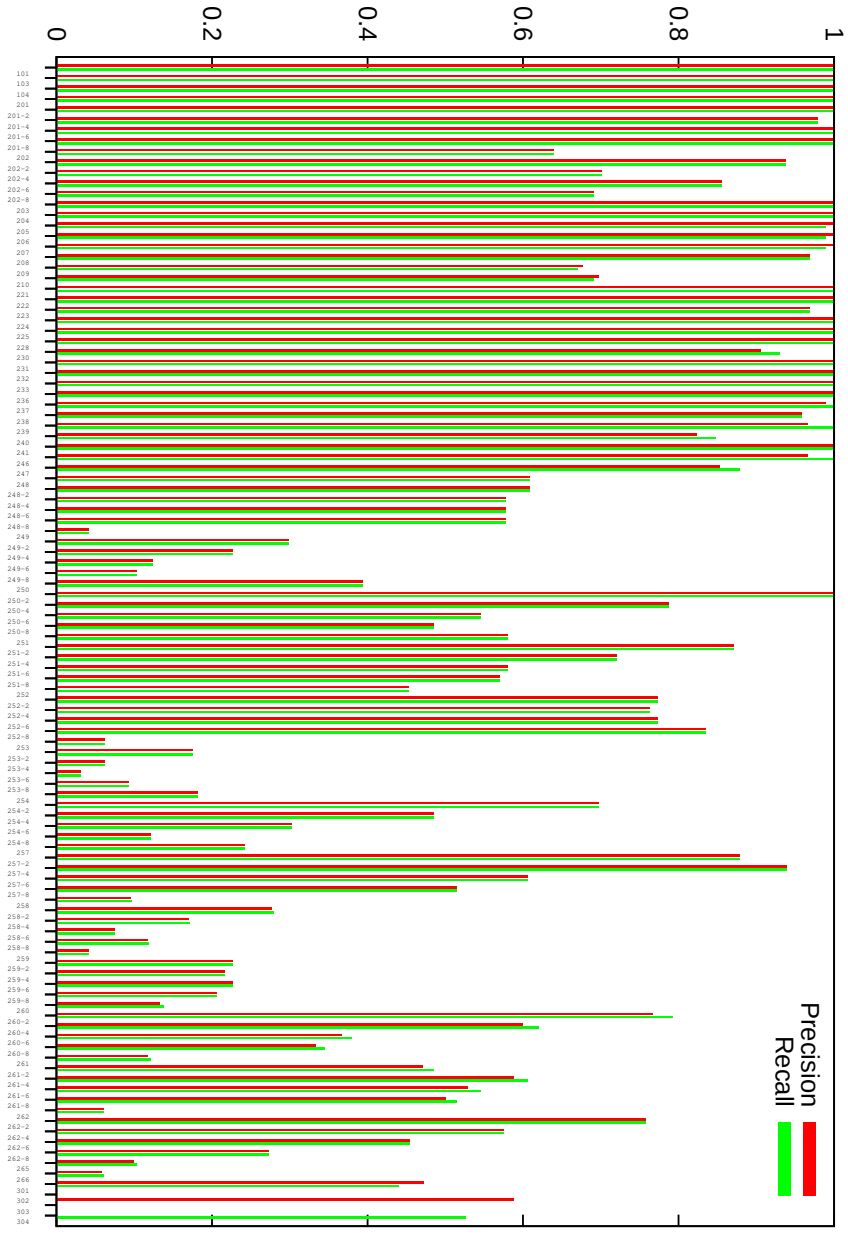

Fig. 1. Results of MapPSO in the OAEI 2009 benchmark data set.

information present in the ontologies. With respect to the quality improvement, it is planned to further investigate in the detailed implementation of these base matchers. In particular, there are plans to incorporate implicit knowledge inferred by a reasoner, as well as more sophisticated graph similarity measures. It is also necessary to review the similarity aggregation for each correspondence in order to better respect the different characteristics of different ontologies by weighting them differently.

There are further plans to deploy the system on a larger computing platform, such as a cloud infrastructure in order to utilise the full potential of the parallel nature of the system. This will be a small step with large impact, as it enables the tool to process large ontologies in reasonable time.

4 Conclusion

The results of the MapPSO system in the benchmark dataset of the OAEI 2009 have been presented. Compared to last year, the system has been extended mainly in terms of additional and refined base matchers, as proposed in the future plans section of last year's contribution [1]. This development resulted in a significant improvement of the alignment results. Future developments will focus on the scalability of the system by enabling the full potential of the parallel nature of the algorithm.

References

1. Bock, J., Hettenhausen, J.: MapPSO Results for OAEI 2008. In Shvaiko, P., Euzenat, J., Giunchiglia, F., Stuckenschmidt, H., eds.: Proceedings of the 3rd International Workshop on Ontology Matching (OM-2008). Volume 431 of CEUR Workshop Series., CEUR-WS.org (November 2008)
2. Correa, E.S., Freitas, A.A., Johnson, C.G.: A New Discrete Particle Swarm Algorithm Applied to Attribute Selection in a Bioinformatics Data Set. In: Proceedings of the 8th Genetic and Evolutionary Computation Conference (GECCO-2006), New York, NY, USA, ACM (2006) 35–42
3. Correa, E.S., Freitas, A.A., Johnson, C.G.: Particle Swarm and Bayesian Networks Applied to Attribute Selection for Protein Functional Classification. In: Proceedings of the 9th Genetic and Evolutionary Computation Conference (GECCO-2007), New York, NY, USA, ACM (2007) 2651–2658
4. Stoilos, G., Stamou, G., Kollias, S.: A String Metric For Ontology Alignment. In Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A., eds.: Proceedings of the 4rd International Semantic Web Conference (ISWC). Volume 3729 of LNCS., Berlin, Springer (November 2005) 624–637
5. Salton, G., Wong, A., Yang, C.S.: A Vector Space Model for Automatic Indexing. Communications of the ACM **18**(11) (1975) 613–620