OntoDNA: Ontology Alignment Results for OAEI 2007

Ching-Chieh Kiu[1], Chien–Sing Lee[2]

Faculty of Information Technology, Multimedia University,
Jalan Multimedia, 63100 Cyberjaya, Selangor. Malaysia.
[1] cckiu@mmu.edu.my, [2] cslee@mmu.edu.my

Abstract. OntoDNA is an automated ontology mapping and merging system that utilizes unsupervised data mining methods, comprising of Formal Concept analysis (FCA), Self-Organizing map (SOM) and K-means incorporated with lexical similarity, namely Levenshtein edit distance. The unsupervised data mining methods are used to resolve structural and semantic heterogeneities between ontologies, meanwhile lexical similarity is used to resolve lexical heterogeneity between ontologies. OntoDNA generates a merged ontology in concept lattice that enables visualization of the concept space based on formal context. This paper briefly describes the OntoDNA system and discusses the obtained alignment results on some of the OAEI 2007 dataset. The paper also presents strengths and weaknesses of our system and the method to improve the current approach.

1 Presentation of the system

1.1 State, purpose, general statement

OntoDNA is an automated ontology mapping and merging tool that provides a scalable environment for interoperating ontologies between information sources. OntoDNA aims to offer contextual and robust ontology mapping and merging through hybrid unsupervised clustering techniques, which comprises of Formal Concept Analysis (FCA) [1], Self-Organizing Map (SOM) and K-Means clustering [2] incorporated with a lexical measurement, Levenshtein edit distance [3]. OntoDNA generates a merged ontology in concept lattice form that enables visualization of the concept space based on formal context.

1.2 Specific techniques used

Ontology is formalized as a tuple $O_i = (C, S_C, P, S_P, A, I)$, where C is concepts of ontology and S_C corresponds to the hierarchy of concepts. The relationship between the concepts is defined by properties of ontology, P whereas S_P corresponds to the hierarchy of properties. A refers to axioms used to infer knowledge from existing knowledge and I instances of concept [4]. The OntoDNA resolves heterogeneous

ontologies by capturing ontological concepts (C) and its ontological elements (S_C, P, S_P, A) [5].

The OntoDNA utilizes FCA to capture the properties and the inherent structural relationships among ontological concepts of heterogeneous ontologies. The captured structures of ontological concepts act as background knowledge to resolve semantic interpretations in similar (synonymy) or different contexts (polysemy).

The unsupervised clustering techniques, Self-Organizing Map (SOM) and K-Means are used to overcome the absence of prior knowledge to discover the structural and semantic heterogeneities between ontologies. SOM organizes ontological elements, clustering more similar ontological concepts together. The clusters of the ontological concepts are derived from the natural characteristics of the ontological elements. Meanwhile K-Means is used to reduce the problem size of the SOM map for efficient semantic heterogeneous discovery in different contexts.

The OntoDNA relies on lexical similarity to resolve lexical heterogeneity by both ontological concept and property names. The lexical similarity, Levenshtein edit distance with the threshold value 0.8 [5] is applied to discover lexical similarity. Prior to the discovery of the degree of lexical similarity, linguistic processing such as case normalization, blank normalization, digit normalization, namespace prefixes elimination, link stripping, and stopword filtering are applied to normalize ontological elements.

The OntoDNA automated ontology mapping and merging framework is depicted in Figure 1. The terms used in the OntoDNA framework are defined as follows:

- Source ontology O_S: Source ontology is the local data repository ontology
- Target ontology O_T: Target ontology refers to non-local data repository ontology
- Formal context K_S and K_T: Formal context K_S is the formal context representation of the conceptual relationship of the source ontology O_S, meanwhile formal context K_T is the formal context representation of the conceptual relationship of the target ontology O_T.
- Reconciled formal context ${}_RK_S$ and ${}_RK_T$: Reconciled formal context ${}_RK_S$ and ${}_RK_T$ are formal context with normalized intents of source and target ontological concepts' properties.
- The ontological elements $O := (C, S_C, P, S_P, A)$: C is concepts of ontology and S_C corresponds to the hierarchy of concepts. P is properties of ontology, and S_P corresponds to the hierarchy of properties. A refers to axioms.

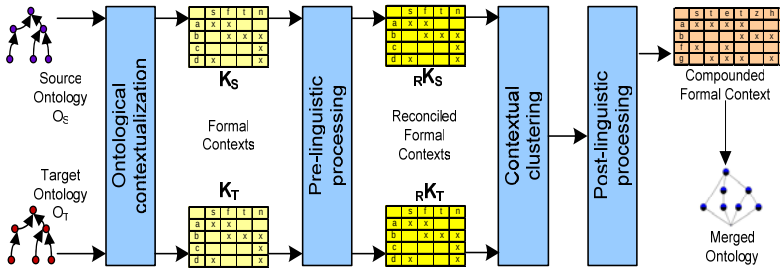

Figure 1. OntoDNA's framework

The OntoDNA algorithmic framework implementation of the automated mapping and merging process illustrated in Figure 1 is explicated below [5] [6]:

Input : Two ontologies that are to be merged, O_S (source ontology) and O_T (target ontology).

Step 1 : Ontological contextualization

The conceptual pattern of O_S and O_T is discovered using FCA. Given an ontology $O := (C, S_C, P, S_P, A)$, O_S and O_T are contextualized using FCA with respect to the formal context, K_S and K_T. The ontological concepts C are denoted as G (objects) and the rest of the ontology elements, S_C, P, S_P and A are denoted as M (attributes). The binary relation $I \subseteq G \times M$ of the formal context denotes the ontology elements, S_C, P, S_P and A corresponding to the ontological concepts C.

Step 2 : Pre-linguistic processing

String normalizations are applied to transform attributes in K_S and K_T prior to lexical similarity mapping. The mapping rules (Map_Rule 1 and Map_Rule 2) (Table 1) are applied to reconcile intents in K_S and K_T. The reconciled formal context ${}_RK_S$ and ${}_RK_T$ are output as input for semantic similarity discovery in the next step.

Step 3 : Contextual clustering

SOM and k-means are applied for semantic similarity mapping based on the conceptual pattern discovered in the formal context. First, the formal context ${}_RK_T$ is trained by SOM. This is followed by k-means clustering to reduce the problem size of the SOM clusters as validated by the Davies-Bouldin index. Subsequently, the formal concepts ${}_RK_S$ are fed to the trained SOM. The source ontological concepts are assigned to the same cluster as their Best Matching Units (BMUs) in the target ontology.

Step 4 : Post-linguistic processing

The mapping rules (Map_Rule 1 and Map_Rule 2) (Table 1) are applied to discover semantic similarity between ontological concepts in the clusters. The ontological concepts of the target ontology are updated to the source ontology based on merging rules (Merge_Rule 1 and Merge_Rule 2) (Table 1).

Output : Merged ontology in a concept lattice is formed.

Mapping Rules
Given source ontological element $O_{elementSi}$ and target ontological element $O_{elementTj}$, apply lexical similarity measure (*LSM*) to map the target ontology O_T to the source ontology O_S at threshold value, *t*, where elements *i* and *j* = 1, 2, 3, …, n. ***Map_Rule 1:*** map ($O_{elementTj} \rightarrow O_{elementSi}$), if LSM($O_{elementSi}$, $O_{elementTj}$) ≥ t; the target ontological element, $O_{elementTj}$ is mapped to (integrated with) the source ontological element, $O_{elementSi}$ and the naming convention and structure of the source ontological element, $O_{elementSi}$ are preserved. ***Map_Rule 2:*** merge ($O_{elementTj} \rightarrow O_S$), if LSM($O_{elementSi}$, $O_{elementTj}$) < t; the target ontological element, $O_{elementTj}$ is merged (appended) to the source ontology and the naming convention and structure of the target ontological element, $O_{elementTj}$ are preserved.
Merging Rules
Given the source ontology O_S in a reconciled formal context $k = (G, M, I)$ and target ontology O_T in a reconciled formal context $l = (H, N, J)$. The source ontology is the base for ontology merging. ***Merge_Rule 1:*** If Map_Rule 1 or Map_Rule 3 is true, the intents of $O_{elementTj}$ (ontological concepts) and its object-attribute relationship $J \subseteq H \times N$ is aligned (appended) into formal context *k*. ***Merge_Rule 2:*** If Map_Rule 2 is true, and formal context *k* is defined by ($O_{extentS1}$, $O_{intentS1}$) ≤ ($O_{extentS2}$, $O_{intentS2}$) :⇔ $O_{extentS1} \subseteq O_{extentS2}$ (⇔$O_{intentS1} \subseteq O_{intentS2}$) the intents of $O_{elementTj}$, its object-attribute relationship $J \subseteq H \times N$ and its *subconcept - superconcept relation* of $O_{elementTj}$ among other concepts are aligned into formal context *k*, whereas the structural relationships of the appended concept is updated with the target ontology as the base.

Table 1. Ontology mapping and merging rules

1.3 Adaptations made for the evaluation

There is no special adaptation for the tests in the Ontology Alignment Evaluation Initiative (OAEI) 2007 campaign. However, a small program is written to translate our native alignment format in the form that is required by the OAEI contest. The URI for benchmark ontology 302 has been manually replaced in order to output the alignment file.

1.4 Link to the system, parameters file and to the set of provided alignments

The OntoDNA system and the alignment results in a ZIP file organized as presented can be downloaded from http://pesona.mmu.edu.my/~cckiu/OAEI2007.htm.

2 Results

The OAEI 2007 campaign provides four ontology tracks, which consist of benchmark, anatomy, directories and thesauri and conference. Due to the ontologies' file size, we manage only to run the alignment tests on the benchmark, directories and conference tracks. In this section, we discuss the results on the benchmark track followed by the experimental outcomes on other tracks.

2.1 Comparison track: benchmark

The benchmark track consists of 51 alignment tests. The alignment results can be divided into five categories for discussion, i.e. Tests 101 – 104, Tests 201 – 210, Tests 221 – 247, Tests 248 – 266 and 301 – 304. The full result of all the alignment tests can be referred in the Appendix.

Tests 101 – 104: The alignment tests consist of the reference alignment, irrelevant ontology, language generalization and language restriction. Overall performance of OntoDNA in the tests is good. The OntoDNA has no problem handling the language generalization (test 103) and language restriction (test 104) features in the tests. The average precision and recall achieved by the OntoDNA are 0.94 and 1.00 respectively as shown in Table 2.

Test	Name	Prec.	Rec.	Time (sec)
101	Reference alignment	0.94	1.00	6.53
102	Irrelevat ontology	NaN	NaN	169.83
103	Language generalization	0.94	1.00	6.36
104	Language restriction	0.94	1.00	6.14
	Average	0.94	1.00	47.22

Table 2. Alignment result for Tests 101 – 104

Tests 201 – 210: The alignment tests manipulate names and comments. Since the OntoDNA relies on the name of classes and properties to resolve lexical heterogeneity, this has resulted in very poor performance in terms of precision and recall for tests 201 and 202 as the name of the labels are not provided. The alignment results on tests 206, 207 and 210 are also poor as the name of the labels are in French translations, and OntoDNA does not understand non-English translations. In addition, as the OntoDNA does not use any thesaurus for resolving lexical similarity, it can't perform well in tests 205 and 209 as illustrated in Table 3.

Test	Name	Prec.	Rec.	Time (sec)
201	No names	0.11	0.01	9.77
202	No names, no comments	0.11	0.11	9.13
203	No comments	0.94	1.00	6.17
204	Naming conventions	0.93	0.84	8.25
205	Synonyms	0.57	0.12	9.31
206	Translation (name)	0.69	0.23	8.61
207	Translation (name and comments)	0.69	0.23	8.52
208	Naming conventions, no comments	0.93	0.84	7.05
209	Synonyms, no comments	0.57	0.12	8.72
210	Translation, no comments	0.69	0.23	8.45
	Average	0.62	0.37	8.40

Table 3. Alignment result for Tests 201 – 210

Tests 221 – 247: The alignment tests manipulate hierarchy. The overall performance of the OntoDNA is good with any kind of hierarchy manipulation (no specialization, flattened hierarchy and expanded hierarchy). However, the OntoDNA alignment results for tests 228, 233, 236, 239, 240, 241, 246 and 247 are poor when the properties are suppressed from the tests as displayed in Table 4.

Test	Name	Prec.	Rec.	Time (sec)
221	No specialisation	0.93	0.76	6.38
222	Flatenned hierachy	0.94	1	7.69
223	Expanded hierarchy	0.94	1	8.69
224	No instance	0.94	1	6.16
225	No restrictions	0.94	1	6.14
228	No properties	0.53	0.27	4.95
230	Flatenned classes	0.91	1	5.97
231	Expanded classes	0.94	1	6.50
232	No specialisation, no instance	0.93	0.76	6.42
233	No specialisation, no properties	0.53	0.27	4.97
236	No instance, no properties	0.53	0.27	4.89
237	Flatenned hierachy, no instance	0.94	1	5.94
238	Expanded hierachy, no instance	0.94	1	8.61
239	Flatenned hierachy, no properties	0.5	0.31	4.94
240	Expanded hierachy, no properties	0.5	0.27	7.06
241	No specialisation, no instance, no properties	0.53	0.27	5.44
246	Flatenned hierachy, no instance, no properties	0.5	0.31	5.03
247	Expanded hierachy, no instance, no properties	0.5	0.27	6.86
	Average	0.75	0.65	6.26

Table 4. Alignment result for Tests 221 - 247

Tests 248 – 266: The alignment tests manipulate hierarchy, labels and comments. The precision and recall of the tests achieved by the OntoDNA are very poor as the names and properties are suppressed as shown in Table 5. The results have proven

that the OntoDNA ~~is~~ strictly relies on ontological concepts and properties name for mapping and merging the ontologies.

Test	Name	Prec.	Rec.	Time (sec)
248	No names, no comments, no specialisation	0.11	0.01	9.23
249	No names, no comments, no instance	0.11	0.01	9.23
250	No names, no comments, no properties	0	0	5.95
251	No names, no comments, flatenned hierachy	0.11	0.01	8.89
252	No names, no comments, expanded hierachy	0.11	0.01	12.66
253	No names, no comments, no specialization, no instance	0.11	0.01	9.28
254	No names, no comments, no specialization, no properties	0	0	6.30
257	No names, no comments, no instance, no properties	0	0	5.91
258	No names, no comments, flatenned hierachy, no instance	0.11	0.01	9.95
259	No names, no comments, expanded hierachy, no instance	0.11	0.01	13.2
260	No names, no comments, flatenned hierachy, no properties	0	0	5.86
261	No names, no comments, expanded hierachy, no properties	0	0	8.13
262	No names, no comments, no specialization, no instance, no properties	0	0	6.09
265	No names, no comments, flatenned hierachy, no instance, no properties	0	0	5.75
266	No names, no comments, expanded hierachy, no instance, no properties	0	0	7.88
	Average	0.05	0.00	8.29

Table 5. Alignment result for Tests 248 - 266

Tests 301 – 304: The alignment tests consist of real bibliographic ontologies. The average precision and recall on the tests are 0.90 and 0.69 respectively achieved by the OntoDNA (Table 6). The results in the tests show that the OntoDNA is a viable automated ontology mapping and merging tool to resolve the heterogeneity of the real ontologies from disparate information sources.

Test	Name	Prec.	Rec.	Time (sec)
301	BibTeX/MIT	0.88	0.69	5.84
302	BibTeX/UMBC	0.9	0.4	5.53
303	Karlsruhe	0.9	0.78	9.95
304	INRIA	0.92	0.88	6.77
	Average	0.90	0.69	7.02

Table 6. Alignment result for Tests 301 - 304

2.2 Expressive ontology: anatomy

We are not able to perform the alignment test on this ontology track due to the large size of the ontology files.

2.3 Directories and thesauri

In this ontology track, there are four ontology alignment tests, i.e., directory, food, environment and library. We manage only to perform the alignment test on the directory ontologies. The alignment tests are not run on food, environment and library ontologies due to the large size of the ontology files.

The directory track is the real world case of websites directory consisting of 4640 alignment tests. Each of the alignment tests contains source and target ontologies. The ontologies are taxonomic ontologies as each of the ontologies contains only classes with superclass-subclass relationships. Since the organizers do not provide the alignment results, we expect feedback on the OntoDNA performance on the directory alignment tests from the organizers.

2.4 Consensus workshop: conference

The conference track consists of 14 real conference ontologies from conference organizations. We have performed 182 alignment tests by aligning an ontology to other ontologies (14 x 13) in the track. Since the alignment results are not provided by the organizers, we expect feedback on the OntoDNA performance on the conference alignment tests from the organizers.

3 General comments

In this section, we summarize the strengths and weaknesses of the OntoDNA and discuss the methods to improve the OntoDNA algorithm.

3.1 Comments on the results

The OntoDNA is an automated ontology mapping and merging tool. All the parameters such as threshold value used for the lexical similarity discovery and the clustering parameters used for structural and semantic similarity discovery are predetermined based on the experimental results on numerous datasets [5][6]. Thus the OntoDNA is a viable tool for mapping and merging ontologies without requiring prior knowledge of the source and target ontological elements.

The limitation of the OntoDNA is the system strictly relies on the name of the ontological concepts and properties to resolve the heterogeneity of ontologies. Thus if the given ontology does not contain the name of the ontological concepts and properties, the OntoDNA is not able to discover lexical similarity for resolving the structural and semantic heterogeneous between the source and target ontologies.

However, given the name of the ontological concepts and properties, the tests results have confirmed that the OntoDNA is an effective system for mapping and merging real ontologies without human intervention in the mapping and merging processes.

3.2 Discussions on the way to improve the proposed system

Based on the tests results, the OntoDNA may need to consider other ontological elements as core elements for mapping and merging. Thus, the structural approach and logic approach can be extended into the OntoDNA algorithm to discover the alignment between source and target ontologies when the ontological concepts and properties name are suppressed (absent). A multi-strategy approach combining linguistic, structural and logic approaches with specific threshold value might also improve OntoDNA's performance.

4 Conclusion

The participation in the OAEI 2007 campaign enables us to identify the strengths and weaknesses of the OntoDNA algorithm and also the methods to improve the OntoDNA algorithm. The presented alignment results show that the OntoDNA has performed well in both ontological concept and property names for mapping and merging ontologies automatically.

References

1. Ganter, B., Wille, R.: Applied Lattice Theory: Formal Concept Analysis, http://www.math.tudresden.de/~ganter/psfiles/concept.ps, (1997).
2. Vesanto, J., Alhoniemi, E.: Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks, vol. 11(3), (2000) 586-600.
3. Do, H., Melnik, S., Rahm, E.: Comparison of schema matching evaluations. In Proc. of the second int. workshop on Web Databases (German Informatics Society) (2002).
4. Gruber, T.R.: A Translation Approach to portable Ontology Specifications. Knowledge Acuquisition, vol. 5, (1993) 199-220.
5. Kiu, C. C., Lee. C. S.: Ontology Mapping and Merging through OntoDNA for Learning Object Reusability. Educational Technology & Society, 9(3), (2006) 27-42.
6. Lee. C. S., Kiu, C. C.: A concept-based graphical – neural approach to ontological interoperability. WSEAS Transactions on Information Science and Applications, 2 (6), (2005) 761-770.

Appendix: Raw results

All the alignment tests are carried out using a notebook with Core Duo T2250 1.73 GHz processor and 1GB RAM in Window XP environment. The precision and recall on the alignment tests with machine processing time in hh.mm.ss.mms format are presented here.

Matrix of results

#	Name	Prec.	Rec.	Time
101	Reference alignment	0.94	1.00	00:00:06:53
102	Irrelevat ontology	NaN	NaN	00:02:49:83
103	Language generalization	0.94	1.00	00:00:06:36
104	Language restriction	0.94	1.00	00:00:06:14
201	No names	0.11	0.01	00:00:09:77
202	No names, no comments	0.11	0.11	00:00:09:13
203	No comments	0.94	1.00	00:00:06:17
204	Naming conventions	0.93	0.84	00:00:08:25
205	Synonyms	0.57	0.12	00:00:09:31
206	Translation (name)	0.69	0.23	00:00:08:61
207	Translation (name and comments)	0.69	0.23	00:00:08:52
208	Naming conventions, no comments	0.93	0.84	00:00:07:05
209	Synonyms, no comments	0.57	0.12	00:00:08:72
210	Translation, no comments	0.69	0.23	00:00:08:45
221	No specialisation	0.93	0.76	00:00:06:38
222	Flatenned hierachy	0.94	1.00	00:00:07:69
223	Expanded hierarchy	0.94	1.00	00:00:08:69
224	No instance	0.94	1.00	00:00:06:16
225	No restrictions	0.94	1.00	00:00:06:14
228	No properties	0.53	0.27	00:00:04:95
230	Flatenned classes	0.91	1.00	00:00:05:97
231	Expanded classes	0.94	1.00	00:00:06:50
232	No specialisation, no instance	0.93	0.76	00:00:06:42
233	No specialisation, no properties	0.53	0.27	00:00:04:97
236	No instance, no properties	0.53	0.27	00:00:04:89
237	Flatenned hierachy, no instance	0.94	1.00	00:00:05:94
238	Expanded hierachy, no instance	0.94	1.00	00:00:08:61
239	Flatenned hierachy, no properties	0.50	0.31	00:00:04:94
240	Expanded hierachy, no properties	0.50	0.27	00:00:07:06
241	No specialisation, no instance, no properties	0.53	0.27	00:00:05:44
246	Flatenned hierachy, no instance, no properties	0.50	0.31	00:00:05:03
247	Expanded hierachy, no instance, no properties	0.50	0.27	00:00:06:86
248	No names, no comments, no specialisation	0.11	0.01	00:00:09:23
249	No names, no comments, no instance	0.11	0.01	00:00:09:23
250	No names, no comments, no properties	0.00	0.00	00:00:05:95
251	No names, no comments, flatenned hierachy	0.11	0.01	00:00:08:89
252	No names, no comments, expanded hierachy	0.11	0.01	00:00:12:66
253	No names, no comments, no specialization, no instance	0.11	0.01	00:00:09:28
254	No names, no comments, no specialization, no properties	0.00	0.00	00:00:06:30
257	No names, no comments, no instance, no properties	0.00	0.00	00:00:05:91
258	No names, no comments, flatenned hierachy, no instance	0.11	0.01	00:00:09:95
259	No names, no comments, expanded hierachy, no instance	0.11	0.01	00:00:13:20
260	No names, no comments, flatenned hierachy, no properties	0.00	0.00	00:00:05:86
261	No names, no comments, expanded hierachy, no properties	0.00	0.00	00:00:08:13
262	No names, no comments, no specialization, no instance, no properties	0.00	0.00	00:00:06:09
265	No names, no comments, flatenned hierachy, no instance, no properties	0.00	0.00	00:00:05:75
266	No names, no comments, expanded hierachy, no instance, no properties	0.00	0.00	00:00:07:88
301	BibTeX/MIT	0.88	0.69	00:00:05:84
302	BibTeX/UMBC	0.90	0.40	00:00:05:53
303	Karlsruhe	0.90	0.78	00:00:09:95
304	INRIA	0.92	0.88	00:00:06:77