Optima+ Results for OAEI 2012

Uthayasanker Thayasivam, Tejas Chaudhari, and Prashant Doshi

THINC Lab, Department of Computer Science, University of Georgia, Athens, Georgia 30602
{uthayasa,tejas,pdoshi}@cs.uga.edu

Abstract. In this report, we present the results of Optima+ in the Ontology Alignment Evaluation Initiative (OAEI) 2012. We mainly foucused on three tracks Benchmark, Conference, and Anatomy. However we were eavluated in all the tracks of the campaign offered in SEALS platform: Benchmark, Conference, Anatomy, Multifarm, Library, and LargeBioMed. We present the new and improved implementation of the Optima algorithm, Optima+ and its results for all the tracks offered within SEALS platform. Optima+ is the latest version of Optima , aimed to perform faster and better. Importantly, we match the highest f-measure (0.65) obtained for the conference track in last year's campaign. Moreover, this year we debut in large ontology tracks: Anatomy and Library aided by a naive divide and conquer approach.

1 Presentation of the system

The increasing popularity and utility of the semantic web increase the number of ontologies in the web. The applications such as web service compositions and semantic web search which utilize these ontologies demand a means to align these ontologies. At present we witness numerous ontology alignment algorithm and tools, that includes more than fifty ontology matching tools in SEALS platform [6] and many more which are not yet reported in SEALS platform [12, 2]. They can be broadly identified using their similarity measures, alignment algorithm and alignment extraction technique. We present a fully automatic general purpose ontology alignment tool called Optima+ , a new and improved implementation of its ancestor Optima [4].

Optima alignment process starts by generating a seed alignment using the lexical attributes of concepts (classes and properties) of the given ontology pair. Then it searches the space of candidate alignments in an iterative fashion and finds the best alignment which maximizes the likelihood. This likelihood estimation exploits the heuristic that the chance of a node pair in correspondence increases if their children are already mapped. Optima algorithm utilizes the lexical similarity between nodes within its structural matching such that its algorithm interlaces both structural and lexical attributes of nodes to arrive at an alignment. We brief out the formal model of an ontology as utilized by Optima and the alignment algorithm adopted by Optima in the next two subsections.

1.1 Ontology Model

The ontology alignment problem is to find a set of correspondences between two ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$. Because ontologies may be modeled as labeled graphs (though with some possible loss of information), the problem is often cast as a matching problem between such graphs. An ontology graph, $\mathcal{O}$, is defined as, $\mathcal{O} = \langle V, E, L\rangle$, where V is the set of labeled vertices representing the entities, E is the set of edges representing the relations, which is a set of ordered 2-subsets of V, and L is a mapping from each edge to its label. Let M be the standard $|V_1| \times |V_2|$ matrix that represents the match between the two graphs $\mathcal{O}_\infty = \langle V_1, E_1, L_1\rangle$, $\mathcal{O}_\in = \langle V_2, E_2, L_2\rangle$:

$$M = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1|V_2|} \\ m_{21} & m_{22} & \cdots & m_{2|V_2|} \\ . & . & \cdots & . \\ m_{|V_1|1} & m_{|V_1|2} & \cdots & m_{|V_1||V_2|} \end{bmatrix} \tag{1}$$

Each assignment variable in M is,

$$m_{a\alpha} = \begin{cases} 1 \text{ if } f(x_a) = y_\alpha : x_a \in V_1, y_\alpha \in V_2 \\ 0 \text{ otherwise} \end{cases}$$

Where $f(\cdot)$ represents the correspondence between the two ontology graphs. Consequently, M is a binary matrix representing the match.

1.2 EM-based Alignment Algorithm

Optima formulates the problem of inferring a match between two ontologies as a maximum likelihood problem, and solves it using the technique of expectation-maximization (EM) originally developed by Dempster et al. [3]. It implements the EM algorithm as a two-step process of computing expectation followed by maximization, which is iterated until convergence. The expectation step consists of evaluating the expected log likelihood of the candidate alignment given the previous iteration's alignment:

$$Q(M^i|M^{i-1}) = \sum_{a=1}^{|V_1|}\sum_{\alpha=1}^{|V_2|} Pr(y_\alpha|x_a, M^{i-1}) \times\ logPr(x_a|y_\alpha, M^i)\pi^i_\alpha \tag{2}$$

Where x_a and y_α are the entities of ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$, respectively, and π^i_α is the prior probability of y_α. $Pr(x_a|y_\alpha, M^i)$ is the probability that node x_a is in correspondence with node y_α given the match matrix M^i. The prior probability is computed using the following equation,

$$\pi^i_\alpha = \frac{1}{|V_1|}\sum_{a=1}^{|V_1|} Pr(y_\alpha|x_a, M^{i-1})$$

The generalized maximization step involves finding a match matrix, M^i_*, that improves on the previous one:

$$M^i_* = M^i \in \mathcal{M} :\ Q(M^i|M^{i-1}_*) \geq Q(M^{i-1}_*|M^{i-1}_*) \tag{3}$$

1.3 Optima+

Optima+ is a new and improved redesign of Optima to achieve a better alignment, yet in significantly less time. It adopts the block coordinate descent (BCD) technique for iterative ontology alignment proposed by us [14] to improve the convergence of the iterative process. Briefly, Optima+ is an optimized and efficient implementation of Optima algorithm. The new features Optima+ brings are 1) Block coordinate descent 2) Improved similarity calculation 3) Improved alignment extraction and 4) Large ontology matching. In the following four sub-sections we describe these four features.

Block Coordinate Descent For Optima Optima+ improve its performance by extending the Optima algorithm with the block coordinate descent (BCD) technique proposed in [14]. This technique helps to speed up its convergence. Let S denote a block of coordinates, which is indexed by a non-empty subset of $\{1, 2, \ldots, N\}$. We may define a set of such blocks as, $B = \{S_0, S_1, \ldots, S_C\}$, which is a set of subsets each representing a coordinate block with the constraint that, $S_1 \cup S_2 \cup \ldots \cup S_C = \{1, 2, \ldots, N\}$. Now, in each iteration, Optima+ (BCD enhanced Optima) chooses a block of the match matrix, $M^i_{S_c}$, and its expected log likelihood is estimated. It chooses the blocks in a sequential manner such that all the blocks are iterated in order. Equation 2 is modified to estimate the expected log likelihood of the block of a candidate alignment as:

$$Q_S(M^i_{S_c}|M^{i-1}) = \sum_{a=1}^{|V_{1,c}|} \sum_{\alpha=1}^{|V_2|} Pr(y_\alpha|x_a, M^{i-1}) \times \; logPr(x_a|y_\alpha, M^i_{S_c}) \; \pi^i_{\alpha,c} \qquad (4)$$

Here, $V_{1,c}$ denotes the set of entities of ontology, $\mathcal{O}_1$, participating in the correspondences included in S_c. Notice that the prior probability, $\pi^i_{\alpha,c}$, is modified as well to utilize just $V_{1,c}$ in its calculations.

The generalized maximization step now involves finding a match matrix block, $M^i_{S_c,*}$, that improves on the previous one:

$$M^i_{S_c,*} = M^i_{S_c} \in \mathcal{M}_{S_c} : Q_S(M^i_{S_c,*}|M^{i-1}_*) \geq Q_S(M^{i-1}_{S_c,*}|M^{i-1}_*) \qquad (5)$$

Here, $M^{i-1}_{S_c,*}$ is a part of M^{i-1}_*. At iteration i, the best alignment matrix, M^i_*, is formed by combining the block matrix, $M^i_{S_c,*}$, which improves the Q_S function as defined in Eq. 5 with the remaining from the previous iteration, $M^{i-1}_{\tilde{S}_c,*}$, unchanged.

An important heuristic, which has proven highly successful in ontology alignment, matches parent entities in two ontologies if their respective child entities were previously matched. This motivates grouping together those variables, $m_{a\alpha}$ in M, into a coordinate block such that the x_a participating in the correspondence belong to the same height leading to a partition of M. The height of an ontology node is the length of the shortest path from a leaf node. Let the partition of M into the coordinate blocks be $\{M_{S_0}, M_{S_1}, \ldots, M_{S_C}\}$, where C is the height of the ontology $\mathcal{O}_1$. Thus, each block is a submatrix with as many rows as the number of entities of $\mathcal{O}_1$ at a height and number of columns equal to the number of all entities in $\mathcal{O}_2$. For example, the correspondences between the leaf entities of $\mathcal{O}_1$ and all entities of $\mathcal{O}_2$ will form the block, M_{S_0}.

Similarity measures Similarity has become a classical tool for ontology matching. Similarity measure between ontological concepts such as classes and properties, is commonly a measure in the range of $[0, 1]$ represents how similar the two concepts are. The similarity measures used in the context of ontology matching can be broadly categorized into lexical similarity and structural similarity. Lexical similarity measures use the lexical properties of a concept (URIs, labels, names, and comments) to measure the similarity between the concepts while structural similarity measures exploit the graph matching algorithms to derive the similarity measure. The lexical similarity used in Optima+ between two concepts C_1 and C_2 is defined as,

$$Sim(C_1, C_2) = Max \left\{ \begin{array}{l} SimLex(Label\text{-}C_1, Label\text{-}C_2), \\ SimLex(Name\text{-}C_1, Name\text{-}C_2), \\ Cos(Comment\text{-}C_1, Comment\text{-}C_2) \end{array} \right\} \tag{6}$$

Where $Label\text{-}C_1$,$Name\text{-}C_1$, and $Comment\text{-}C_1$, are the label, name and comment of the concept C_1. As shown in Eq. 7 below the lexical similarity between the phrases P_1 and P_2 is,

$$SimLex(P_1, P_2) = Max \left\{ \begin{array}{l} LinSim(P_1, P_2), CosSim(P_1, P_2), \\ SWSim(P_1, P_2), NWSim(P_1, P_2), \\ LevSim(P1, P2) \end{array} \right\} \tag{7}$$

Here, LinSim is the popular similarity measure introduced by Lin [7] and CosSim is the gloss based cosine similarity described in [15]. These two similarity measures requires a lexical database like WordNet [9]. Optima+ uses WordNet version 3.0 for OAEI 2012 along with the information content database provided by [11]. SWSim is the Smith-Waterman [13] similarity measure and NWSim is the Needleman-Wunsch [10] similarity measure. LevSim is the similarity measure that is the inverse of Levenshtein distance between the phrases.

Alignment Extraction Alignment extraction is the process of pruning a set of correspondences in an alignment to achieve a minimal and consistent alignment. A minimal alignment is achieved by removing the correspondences which can be inferred by an existing correspondence. A consistent alignment is achieved by resolving conflicting correspondences. Optima+ adopts a simple heuristic based alignment extraction process, which is described below,

- For each class-correspondence (N_1, N_2) in the alignment, any correspondence among the children of N_1 and children of N_2 is removed.
- For each class-correspondence (N_1, N_2) in the alignment, any correspondence which maps children of N_1 to parent of N_2 or children of N_2 to parent of N_1 is removed if its similarity is less than the similarity of N_1 and N_2.
- If a concept is mapped to more than one concept then, we select the correspondence with highest similarity (Max_{Sim}) and remove all other correspondences which are less than a predefined threshold T_1. We also remove all other correspondences with similarity less than the $Max_{Sim} - \delta$. Here δ is a user configurable value in the range of $[0, 0.5]$.

Large Ontology Matching The time complexity of Optima to align Ontology O_1 of size $|O_1|$ and O_2 of size $|O_2|$ is $(|O_1| \times |O_2|)^2$ [4]. Hence, despite its efficient implementation in Optima+ , it still takes significantly longer time to match larger ontologies. We solve this problem using a naive divide and conquer approach. The large ontology matching is triggered if number of classes in one of the ontology exceeds a user configurable threshold (for this campaign it is set to 600 named classes). Optima+ partitions the ontology using a structural partitioning algorithm and matches every block from first ontology with every block from the second ontology separately. Finally, it merges all the block-alignments together as final alignment. The partitioning algorithm employed in Optima+ is based on breadth first tree traversal described in [4].

1.4 State, purpose, general statement

Optima+ is a general purpose ontology alignment tool capable of matching English language ontologies described in OWL, RDFS/RDF, and N3.

1.5 Specific techniques used

As described earlier, Optima+ employs a variety of similarity measures, a simple alignment extraction and large ontology matching using a naive divide and conquer approach.

1.6 Adaptations made for the evaluation

We made couple of changes to the alignment extraction process for this campaign. First, we filtered the correspondences between imported concepts even though they have been directly used within the ontologies. Second, we implemented the heuristics mentioned in the sub-section 1.3 to make the alignment minimal. The default alignment extraction of optima is not as strict as the one configured for this campaign.

1.7 Link to the system and parameters file

A detail presentation of the system, its configuration and parameters used for this campaign and results can be found at `http://thinc.cs.uga.edu/thinclabwiki/index.php/OAEI_2012`.

2 Results

Optima+ is evaluated in all the six tracks under SEALS platform in OAEI 2012 though, we only focused in benchmark, conference and anatomy tracks. For this report the results for all these tracks are summarized except for large biomedical track. Optima+ could not successfully finish aligning the large biomedical track due to a fatal error. Detailed results for individual tracks and test cases can be found at `http://thinc.cs.uga.edu/thinclabwiki/index.php/OAEI_2012`.

2.1 benchmark

The Benchmark test library consists of 5 different test suites [8]. Each of the test suits is based on individual ontologies, consists of number of test cases. Each test case discards a number of information from the ontology to evaluate the change in the behavior of the algorithm. There are six categories of such alterations – changing name of entities, suppression or translation of comments, changing hierarchy, suppressing instances, discarding properties with restrictions or suppressing all properties and expanding classes into several classes or vice versa. Suppressing entities and replacing their names with random strings results into scrambled labels of entities. Test cases from 248 till 266 consist of such entities with scrambled labels. Table. 1 shows Optima+ 's performance in benchmark track on, 100 series test cases, 200 series test cases without scrambled labels test cases and all the scrambled labels test cases. The average precision for Optima+ is 0.95 while average recall is 0.83 for all the test cases in 200 series except those with scrambled labels. For test cases with scrambled labels, the average recall is dropped by 0.53 while precision is dropped only by 0.04. When labels are scrambled, lexical similarity becomes ineffective. For Optima+ algorithm, structural similarity stems from lexical similarity hence scrambling the labels makes the alignment more challenging for Optima+ . Result is 46% decrease in average F-Measure from 0.85 to 0.46. This trend of reduction in precision, recall and f-measure can be observed throughout the benchmark track. For all the test suits, test cases with scrambled labels resulted into lower precision, recall and f-measure. Optima+ 's algorithm faces difficulties in aligning ontologies with low or no lexical similarity.

	Bibliography			2			3			4			Finance		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
100 Series	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
201-247	0.88	0.85	0.85	1	0.84	0.87	0.97	0.88	0.89	0.93	0.77	0.79	0.96	0.8	0.83
248-266	0.65	0.35	0.43	1	0.36	0.46	0.98	0.38	0.49	0.96	0.34	0.43	0.96	0.38	0.49

Table 1. Performance of Optima+ in OAEI 2012 for benchmark track

2.2 anatomy

Previous year, Optima could not sucessfully complete aliging anatomy track. This year, with the help of large ontology matching process, Optima+ is able to sucessfully align ontologies of this track. In anatomy track, Optima+ yields 0.854 precision and 0.584 recall in 6460 seconds. We hope with bio medical lexical databases like Unified Medical Language System (UMLS) [1] Optima+ could improve its recall.

2.3 conference

For this track, Optima+ achieves recall of 0.68 and precision of 0.62. Both the recall and the precision are improved compared to the performance of Optima in OAEI 2011. Overall there is 81% increase in F-Measure compared to OAEI 2011. This makes Optima+ , to tie the top performer in OAEI 2011[5] in terms of F-Meaure(0.65). Table 2 lists the harmonic means for precision, recall and f-measure along with total runtime for conference track of Optima in OAEI 2011 and Optima+ in OAEI 2012.

The performance improvement in conference track arises from the improved similarity measure and the alignment extraction (Section 1.3). Optima+ also utilizes improved design and optimization techniques to reduce the runtime. The runtimes reported in the Table 2 cannot be compared directly as the underlying systems used for evaluations differ. However, the runtime improvement from 15+ hours to around 23 minutes is perspicuous.

Year	Precision (H-mean)	Recall (H-mean)	F-Measure (H-mean)	Total Runtime
2011	0.26	0.60	0.36	15hrs
2012	0.62	0.68	0.65	1349sec

Table 2. Comparison between performances of Optima+ in OAEI 2012 and Optima in OAEI 2011 for conference track

2.4 multifarm

Since Optima+ focus only on English language ontologies, it gives low performance in this track as expected. However it is interesting to notice that Optima+ yields an average recall of 1.0 with an average precision of 0.01.

2.5 library

Library is another large ontology matching track in OAEI 2012. Optima+ attains a precision of 0.321 and a recall of 0.072 in 37,457 seconds.

3 General comments

Last year Optima debuted the OAEI campaign with promising results. However it took too long to finalize the alignment process. This year we redesigned the Optima algorithm to complete the alignment process faster and were able to speed it from minutes to seconds. Additionally, we implemented a naive divide and conquer approach to tackle the large ontology matching problem.

Optima+ matches the last year's best f-measure (0.65) in conference track, and gives 0.87 f-measure on average for benchmark track excluding the scrambled labeled test cases. However, as revealed in benchmark track Optima+ heavily relies on lexical features of ontologies to align them. In large ontology tracks (anatomy and library) Optima+ struggles to perform well as it performed in other tracks (conference and benchmark). We suppose that a dedicated alignment extraction is needed to merge the results of blocks in large ontology matching process.

We are aiming to improve our f-measure for large ontology matching by improving the entire large ontology matching process. Specifically, we would like to introduce an exclusive alignment extraction process for large ontology matching. Further, we want to find an optimum partition strategy for BCD technique which yields better alignment yet faster. On top of these, extending the current similarity measure calculation with more useful similarity measures and lexical databases would help Optima+ to improve its f-measure. Though there is an inherent means to align instances using Optima algorithm, Optima+ implementation is not yet fully capable of matching instances. In its next versions, we expect it to be able to match instances as well.

4 Conclusion

In this report we present the results of Optima+ in OAEI 2012 campaign in six tracks including Benchmark, Conference, Anatomy, Multifarm, Library, and LargeBioMed. We also present the new and redesigned implementation of Optima , Optima+ . Optima+ shows impressive performance in benchmark track, but struggles to align ontologies with scrambled labels. However, it matches the top f-measure of last year's conference track. It debuted in large ontology tracks (anatomy and library) with promising results. In future we want to participate in more tracks, especially instance matching tracks. More importantly, we wish to leverage our performance in large ontology tracks to attain a higher f-measure.

References

1. Bethesda. Umls reference manual. `http://www.ncbi.nlm.nih.gov/books/NBK9676/`, 2009.
2. S. Castano, A. Ferrara, and S. Montanelli. Matching ontologies in open networked systems: Techniques and applications. *Journal on Data Semantics (JoDS)*, V, 2005.
3. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society.*, 39:1–38, 1977.
4. P. Doshi, R. Kolli, and C. Thomas. Inexact matching of ontology graphs using expectation-maximization. *Web Semantics*, 7(2):90–106, 2009.
5. Jérôme Euzenat, Alfio Ferrara, and et al. Results of the ontology alignment evaluation initiative 2011. In *Ontology Matching Workshop ISWC*, 2011.
6. Asuncion Góez-Pérez. Seals. `http://www.seals-project.eu/`, 2012.
7. D. Lin. An information-theoretic definition of similarity. In *ICML*, pages 296–304, 1998.
8. Jose Luis. Benchmark test library. `http://oaei.ontologymatching.org/2012/benchmarks/index.html`, 2012.
9. G. A. Miller. Wordnet: A lexical database for english. In *CACM*, pages 39–41, 1995.
10. Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443 – 453, 1970.
11. Ted Pedersen and Siddharth Patwardhan. Wordnet::similarity - measuring the relatedness of concepts. In *AAAI*, pages 1024–1025, 2004.
12. Quentin Reul and Jeff Z. Pan. Kosimap: Ontology alignments results for oaei 2009. In *Ontology Matching Workshop ISWC*, pages –1–1, 2009.
13. T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. In *JMB*, volume 147, pages 195–197, 1981.
14. Uthayasanker Thayasivam and Prashant Doshi. Improved convergence of iterative ontology alignment using block-coordinate descent. In *AAAI*, pages 150–156, 2012.
15. M. Yatskevich and F. Giunchiglia. Element level semantic matching using wordnet. Technical report, University of Trento, 2007.