PARETO OPTIMIZATION FOR MULTIOBJECTIVE MATCHING OF GEOSPATIAL ONTOLOGIES

Ujwala Bharambe[1], S.S Durbha[1] ,
Kuldeep Kurte[1], Nicolas H. Younan[2], Roger L. King[2]
1. Centre of Studies in Resource Engineering, Indian Institute of Technology Bombay(IITB), Powai, Mumbai - 400076, Maharashtra, INDIA
2. Department of Electrical and Computer Engineering , Mississippi State University, Mississippi State, MS 39762-9571, USA
ujwala.bharambe@iitb.ac.in, sdurbha@iitb.ac.in, kuldeep3101988@gmail.com, younan@ece.msstate.edu, rking@engr.msstate.edu

ABSTRACT

Geospatial information is different than conventional information. Harmonization is needed for interoperability and seamless access to data. Ontology matching is an emerging solution to achieve this harmonization. The input data of the Geospatial ontologies vary from the conventional ontologies and hence it is conceptualized in a different manner. There are two major obstacles for geoinformation fusion: heterogeneity and uncertainty. Heterogeneity is more prevalent and uncertainty is an unavoidable entity in geospatial domain.This paper explores a novel multi-objective algorithm for geospatial ontology matching. It uses Pareto ranking to sort the probable solution and derives the pareto front. This pareto front is used further to find the best match.

Index Terms— Interoperability, Ontology Matching, Pareto Ranking, Pareto Front.

1. INTRODUCTION

Interoperability is a major problem in geospatial environment especially for information systems that use geospatial and remotely sensed data. There is an increasing reliance on distributed web-based access to geospatial information; thus there is a need to increase the efficiency and the speed of fusing geographic information from multiple sources. However, semantic heterogeneity is a major problem faced by many geospatial information fusion systems due to the heterogeneity in different geographic information sources which use a variety of concepts and categories that may not be compatible with each other. For example, if there are three different sources capturing the images of a storm or hurricane, each source will be capturing different details. It then becomes challenging to integrate the inputs of all the three sources and give an integrated view for the purpose of analysis. Hence, to overcome this problem there is a need for resolving heterogeneities between different geospatial information sources [1]. Ontology matching is one such approach which enables the reconciliation of heterogeneous semantic representations by mapping concepts that are similar based on structural, lexical, extensional, subsumption relationships etc. In this approach, the input data from the information sources are conceptualized in terms of ontologies and then these ontologies are matched or mapped in order to fuse the two input data sources. An ontology is a formal specification of a shared conceptualization [2] and is one of the preferred ways to standardize semantics of data. It is reusable and can be adopted for many applications.

Ontology matching is a complex problem. One way of solving multi-objective optimization problem is converting it into single objective optimization problem by combining them in such a way so as to produce an aggregate scalar function (using linear combination). However, this approach to solve multi objective optimization problems has several limitations such as:

- it requires a priori knowledge about the relative importance of the objectives, and the limits on the objectives that are converted into constraints
- the aggregated function leads to only one solution; and
- trade-offs between objectives cannot be easily evaluated [8].

This paper focuses on multi objective geospatial ontology matching based on Pareto optimization to achieve semantic interoperability.

This paper focuses on multi objective geospatial ontology matching based on Pareto optimization to achieve semantic interoperability.

.

2. METHODOLOGY

For any two given ontologies A and B, mapping one ontology with another means that each node in Ontology A is matched with a similar node in Ontology B such that their semantics are matching with each other. For better precision there is a need for defining the semantic relationships that can exist between the two related concepts and to develop

an algorithm to discover such similar concepts in the two ontologies.

For matching different ontologies various types of similarity measures are used: terminological, structural, semantic and instance. The best match is the one which maximizes all the similarity measures. However, in practice there may be conflicting values amongst the similarity measures thus impeding the matching process. To resolve this issue, there is a need for optimization. Optimization means finding a solution which cannot be improved any further. Pareto trade off analysis is a popular method where a multi-objective optimization is performed. This does not give a single solution but a set of solutions called pareto set (paretofront) of minimal elements defined by the dominance relation[7][3].

In this paper we have applied the concept of pareto optimization to find the best match between the two input ontologies.

2.1 Multiobjective optimization and Pareto Optimality

Let a Multi-objective Problem be

$$\min F(x) = \left(f_1(x), \ldots., f_q(x)\right) \quad s.t: \ x \in X$$

$X = \{x \in R^n | g_i(x) \leq 0.\ I=1,\ldots,m\}$, $f_j : R^n \rightarrow R$, j=1….,q and $g_i : R^n \rightarrow R$, i=1,….,m

$F_i(x)$ is i^{th} objective function to be minimized; q is the number of objectives.

X is the feasibility region of the problem and its points are called feasible solutions to the problem.

Definition: Given two solutions x,y ∈ X, we say that $x = (x_1,\ldots..,x_q)$ dominates $y = (y_1, \ldots., y_q)$ if and only if $F(x)$ is partially less than $F(y)$, i.e., if and only if $f_i(x) \leq f_i(y)\ \forall i \ \ =1,\ldots\ldots,q$ and $\exists i \in \{1,\ldots\ldots q\}$ with $f_i(x) \leq f_i(y)$

Definition: One solution x ∈ X is said to be *Pareto optimal* or *efficient* with regard to a set $B \subset A$ if and only if there does not exist any x ∈ *B* such that x dominates y [8].

A key concept of Pareto-optimization [3] is the concept of dominance. The potential solutions to a Pareto-optimization problem can be classified into dominated solutions and non-dominated solutions. A solution is a Pareto-optimal solution if it is not dominated by any other feasible solution [4]

In practical multiobjective optimization problem, it is usually impossible find a unique solution that dominates all other solutions [3]. Instead, it is expected that number of pareto optimal solutions can be found and they together form Pareto front. In this Pareto front, increase in one objective in pareto optimal solutions will surely cause decrease in one or more other objectives. The goal of multi objective optimization function is to generate various feasible solutions which are closer to Pareto front from which the best solution could be selected [5].

3. PROPOSED SYSTEM ARCHITECTURE

In Ontology matching, dominance can be the degree of match between two concepts of source and target ontologies. This paper proposes a novel semi-automated method to match geospatial ontologies based on terminological and structure based matching with adaptive fusion using Pareto optimization.

Proposed System Architecture (shown in Figure 1) is layer based and utilizes several matchers in iteration.

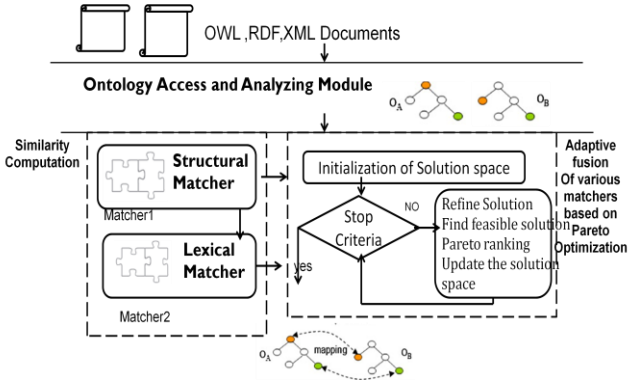

Figure 1: proposed System Architecture

The first layer (Ontology access and analyzing module) takes two input ontologies(OWL/RDF) and performs parsing of the ontologies and some preprocessing tasks. The next layer (Ontology matching layer) is responsible for capturing all lexical (WordNet, Cosine and Jaro Wrinkler) and structure features (KL divergence, KL divergence with joint probability, Cosine Strcture similarity) of the given ontologies. These features are utilized by a variety of matchers for calculation of similarity between two ontologies. We have also used Gaussian based structure similarity measure eq(1) to remove least probable match.

$$gaussvalue_{ij} = e^{\{-d^2/2\sigma^2\}} \quad (1)$$

Where d is Euclidean distance between relative entropy concepts and σ is scaling factor [6].

Algorithm 2: Pareto Ranking

```
Input : size: m and n, MatchingMatrix (MM_ij )
Output : Ranked (MM_ij )
1:   Create List rank_i=-1;
2:   For i=0 to n
3:   {
4:     For j=0 to m
5:     Insert pop_i
6:      Remain=m;
7:     currRank=0;
8:     While(Remain>0)
9:      {
10:       For i=0 to m
11:         If(rank_i == -1)
12:           If(! IsDominate(i ,currRank))
13:           {     rank_i = currRank
14:                   Remain= Remain-1
15:            }
16:         currRank= currRank+1
17:  }}
```

Figure 2: Algorithm for Pareto ranking

Algorithm1 : Ontology Analysis and Matching

Input : Ontology O1 and O2

Output : matched candidates concepts

1: For every concept in ontology $O_{1\,(i)}$ and $O_{2\,(j)}$, Calculate following parameters: p (Structure probability),jp (joint probability),cp (conditional Probability),e(entropy) and H (relative entropy).

2: Calculate lexical similarities (JK_{ij} , WN_{ij}, $CoString_{ij}$) based on lexical features (JK_{ij} : Jaro-Wrinker Similarity, WN_{ij} : Wordnet Noun Similarity, $CoString_{ij}$: Cosine Similarity) LX=min(JK_{ij} , WN_{ij}, $CoString_{ij}$)

3: Calculate structure similarity parameter (KL_{ij}, $CoStructure_{ij}$, $KLJP_{ij}$,)for every concept of between $O_{1\,(i)}$ and $O_{2\,(i)}$ using joint probability distribution(KL: KL divergence CoStructure : Cosine Similarity). ST=min((KL_{ij}, $CoStructure_{ij}$, $KLJP_{ij}$)

4: Initialize Match Matrix(MM_{ij}) with LX and ST

5: Calculate $gvalue_{ij}$ using equation(Gaussian f)

6: If $gvalue_{ij}$ > α

7: Delete average nonfeasible solution from MM_{ij}

8: Iteration Phase(Until Stop criteria are reached)

9: Rank each solution based on domination (highest value in row of Matched Matrix (MM_{ij})

10: Perform Search for feasible solutions (find Pareto front)

11: Update the best found solution.

12: Updates matrix MM_{ij} based on minimum value of LX and ST (updating solution space)

13: Output results: Select the concepts from both ontologies which will have highest Rank.

Figure 3: Algorithm for Ontology Analysis and matching

The matching process is divided into two main modules:

1. Similarity computation composed of various matchers wherein each concept of source ontology is compared with all concepts of the target ontology based on lexical features and structural features (Figure 3)
2. Adaptive fusion of various matchers based on Pareto optimization/ranking (Explained in Figure 2) in which best candidate concepts are selected from input ontologies. For aggregating various matchers, this system takes advantage of Pareto optimization in order to select best matching concept.

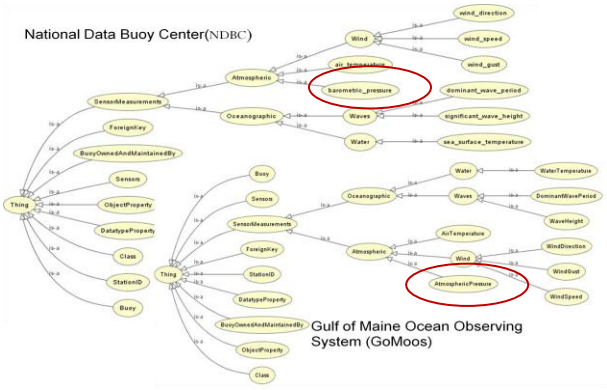

Figure 4: Example of ontologies

4. RESULTS AND DICUSSION

The proposed system is implemented in a Java Programming environment. The system considers two input ontologies (OWL, RDF) which are read using Jena API[10] and the feature parameters (structure and linguistic) are calculated. Various matchers (String: WordNet[12], Cosine and Jaro Wrinkler[11]; Structural: KL divergence[9], Covariance and Gaussian function) are applied on the input ontologies to find the alignment. The output of all these various matchers is stored in a multidimensional matrix. To find the best choice amongst these results, we have used pareto ranking method.

The sample ranking generated for different coastal buoy organization NDBC (National Buoy Data Center), GoMOOS(Gulf of Maine Ocean observation System) ontologies (Figure 4) by ranking and gauss value (eq (1)) is presented in the table no. 1

Table 1
Sample Values obtained from algorithm 1 for NDBC, GoMOOS ontologies

Concept1	Concept2	Structural sim =f1	Lexical sim =f2	Gauss Value	Rank
AtmosphericPressure	barometric_pressure	0.0	0.105	1.0	0
AtmosphericPressure	air_temperature	0.0	0.291	0.9873	0
AtmosphericPressure	Water	0.001	0.485	0.6247	1
AtmosphericPressure	Waves	0.001	0.498	0.4456	1
AtmosphericPressure	significant_wave_height	0.045	0.577	0.4196	2.0
AtmosphericPressure	wind_direction	0.031	0.585	0.4634	2.0

In table 1, if the structural similarity (f1) and lexical similarity is minimum i.e. 0, it suggests a good match while for Gaussian value (eq. no. 1) value of 1 suggests a good match.

As shown in table no 1, the least probable solution candidates are removed from the match pool. In algorithm 1(fig 3) line number 6 represents removal of least probable candidates based on Gaussian value as shown in table no 1. For example, shown in figure 5, is an ontology with 19 concepts which are compared with each of the 19 concepts of other ontology. A total of 361 comparisons are carried out. Ranking these many elements is tedious task. Hence, we have used Gaussian function (eq 1) to remove the least probable solutions. After that the proposed system uses pareto ranking technique to rank the solutions and then the best ranking candidate (0 value) is considered as the best match. Table 2 shows the output for NDBC and Gomoose ontologies

Table 2
Output obtained from algorithm 1 for NDBC, GoMOOS ontologies

Concept1	Concept2	Lexical Similarity	Structure Similarity	Rank	Gauss Value
SensorMeasurements	SensorMeasurements	0.0	0.0	0	1
Oceanographic	Oceanographic Atmospheric	0.0 0.344	0.0 0.0	0	1 0.414
Water	Water	0.0	0.0	0	1.0
WaterTemperature	sea_surface_temperature	0.121	0.0	0	1.0
Waves	Waves Water	0.0 0.213	0.0 0.0	0 0	1.0 0.575
DominantWavePeriod	dominant_wave_period significant_wave_height	0.0 0.379	0.0 0.0	0 0	1.0 0.933
WaveHeight	significant_wave_height dominant_wave_period	0.299 0.505	0.0 0.0	0 0	1.0 0.932
Atmospheric	Atmospheric Oceanographic	0.0 0.344	0.0 0.0	0 0	1.0 0.414
Wind	Wind	0.0	0.0	0	1
WindSpeed	wind_speed wind_direction wind_gust	0.02 0.214 0.178	0.0 0.0 0.0	0 0 0	1 0.96 0.855

This system is tested on various geospatial ontologies NDBC (National Data Buoy Centre), GoMOOS (Gulf of Marine Ocean Observation System), weather ontologies, sensor ontologies and hydrology ontologies. NDBC and GoMOOS ontologies are almost similar, both structurally and terminologically. The BBC, CNN weather service ontologies are also showing a total match in our algorithm [7]. The weather ontologies which are having more than 100 concepts were also matched successfully. Overall matching accuracy of the proposed algorithm is 85%.

5. CONCLUSION

In this paper, we presented a Pareto ranking based ontology matching algorithm for geospatial ontologies. Based on the experiments our algorithm has a good overall performance. Currently, testing has been done on limited set of ontologies; further experiments are required for definitive results. Pareto ranking is a common fitness function used in evolutionary algorithm. Hence, in future we intend to consider evolutionary technique for improving the accuracy.

6. REFERENCES

[1] Buccella Agustina, Alejandra Cechich , Pablo Fillottrani "Ontology-driven geographic information integration: A survey of current approaches ",*Computers & Geosciences* Volume 35, Issue 4, April 2009, Pages 710–723.

[1] Gruber T.R,"A translation approach to portable ontologies", *Knowledge Acquisition*, 5(2):199-220,1993.

[2] Nan Liu, Bo Huang, and Xiaohong Pan,"Using the Ant Algorithm to Derive Pareto Fronts for Multiobjective Siting of Emergency Service Facilities",Transportation Research Record*: Journal of the Transportation Research Board,No. 1935,* Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 120–129.

[3] Lampinen Jouni, "Multiobjective Nonlinear Pareto-Optimization. A Pre-Investigation Report ",[Online]. Available: http://www2.it.lut.fi/kurssit/04-05/010778000/Pareto.pdf

[4] Xiao, N., D. A. Bennett, and M. P. Armstrong. "Using Evolutionary Algorithms to Generate Alternatives for Multiobjective Site-Search Problems". *Environment and Planning ,* Vol. 34, 2002, pp. 639–656.

[5] Weather Ontologies :http://www.scs.ryerson.ca/~bgajdero
CNN and BBC Service Ontologies
http://www.daml.org/services/owl-s

[6] Kunegis, J. DAI-Labor , Lommatzsch, A., Bauckhage, C.,"Alternative Similarity Functions for Graph Kernels",*19th International Conference on Pattern Recognition (ICPR 2008),* December 8-11, 2008, Tampa, Florida, USA. IEEE 2008,Page(s):1-4

[7] Patrick Ngatchou, Anahita Zarei and M.A. El-Sharkawi," Pareto Multi Objective Optimization", in Proceedings of the 13th International Conference on Intelligent Systems Application to Power Systems (ISAP 2005), pp. 84-91, IEEE Press, Washington, DC, USA, 6-10 November, 2005.

[8] Alberto, I.; Azcarate, C.; Mallor, F. & Mateo, P.M.," Multiobjective Evolutionary Algorithms. Pareto Rankings." Monografıas del Semin. Matem. Garcıa de Galdeano 27: 27-35 2003.

[9] Erik P. Blasch, Éric. Dorion., Pierre Valin, Eloi Bossé,"Ontology Alignment using Relative Entropy for Semantic Uncertainty Analysis"., Aerospace and Electronics Conference (NAECON), Proceedings of the IEEE 2010 National Conference, 140 - 148. 14-16 July 2010.

[10] http://jena.sourceforge.net/ontology/index.html

[11] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A comparison of string distance metrics for namematching tasks, 2003.

[12] Giuseppe Pirrò, Jérôme Euzenat: "A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness". Proceedings of the 9th International Semantic Web Conference (ISWC2010). LNCS 6496, Springer 2010, pp. 615-630.