International Journal of Pure and Applied Mathematics
Volume 117 No. 21 2017, 287-299
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Automated Ensemble Framework for Integration of Ontology Based large Scale Semantic Knowledge Base

[1]G. Silambarasan and [2]J. Anvar Shathik
[1]CMJ University,
Meghalaya, India.
gssilambarasan@gmail.com
[2]KGISL Institute of Technology,
Coimbatore, India.
anvarshathik@gmail.com

Abstract

Knowledge base is growing exponentially nowadays using different techniques. The ontology has been used widely to integrate the knowledge base for easy retrieval of the web document contents to the user queries. Several steps have been taken in the literatures to integrate the knowledge base which contains the overlapping and complementary information. In this paper, we propose a novel technique to knowledge based integration named "Automatic Ensemble framework for Integration of Ontology Based Sematic Knowledge Base". It considers the semantic heterogeneous class structures. The proposed framework provides the Solution to the NP hard problem in terms of query selection. Ensemble framework produces the multiple class structures to the knowledge base as knowledge base is large in size and structure matching model is leverages to identify the relationship based on semantic in order to integrate the complex structures of the different KBs. Integrated Knowledge base is been available to access through queries but improper information selection to query leads to complex problem which can be avoided by placing the adaptive query selection algorithm using greedy algorithms . The experimental result demonstrates that proposed model outperforms the state of art approaches in terms of effectiveness, efficiency and accuracy.

Key Words: Ontology, knowledge base, semantic web, data integration, ensemble technique.

1. Introduction

With development of semantic web in recent years more and more data has been published in Semantic Web formats the Resource Description Framework (RDF) and the Web Ontology Language (OWL). Currently large scale knowledge bases (KBs) have been constructed using different technique and from different sources and it becoming large such as YAGO, Probase, Freebase, DBpedia, NELL and DeepDive [1][2]. Mostly KB designed using different technique usually contains overlapping and complementary information. Moreover, as knowledge acquisition is an expensive process, reusing existing KBs is strongly desirable to reduce the cost of data management. Therefore, knowledge base integration has attracted growing interests. In the last decade, a wide variety of works have been conducted on ontology integration [3], [4] which is related to the problem of knowledge base integration, as an ontology can be treated as the conceptual system to underlie a particular knowledge base. To integrate KBs, both data and structure information are combined to align classes, instances and relations/properties. The alignment process has to be found based on class equivalence. Major task in the KB Integration are class structure integration and instance matching. In this work, class structure is represented as Taxonomy. The Class Structure Integration and instance matching is used for entity resolution, data integration and data cleaning.

In this paper. We propose Automatic Ensemble framework for Integration of Ontology Based Sematic Knowledge Base. In this taxonomy integration based on ensemble mechanism is carried out initially and then aligning of the instances is carried out based on the taxonomy integration result. Instance matching is computed with each class structure which is partitioned by ensemble process, ensemble process is proposed in order to reduce the computation time by partitioning the data into different partitions. The relationship between the data is classified into more categories by employing the unsupervised classification model such as Principle component analysis. It is can be used to determine the equivalence relationship and generalization relationship to integrate the two KB to generate the Unified Structure. The remainder of the paper is organized as follows: Section 2 discusses the related works in data integration of KB and its impacts against the performing classification under context information, Section 3 briefly discusses the proposed technique in terms of class structure integration and instance mapping automatically and Section 4 presents the experimental results on a number of data sets. Section 5 discusses conclusions and future work.

2. Related Works

There exist many techniques to integrate the KBs are designed and implemented efficiently. Each of these techniques follows some sort of class structure unification, among few performs nearly equivalent to the proposed framework, which is described as follows.

Actively Learning Ontology Matching Via User Interaction

In this literature, we analyse the active learning framework for ontology matching which tries to find the most informative candidate matches to query of the user.

The user's feedbacks are used to: correct the mistake matching propagates the supervise information to help the entire matching process. Different measures are utilized to estimate the confidence of each matching candidate. A propagation algorithm is further enabled to maximize the spread of the user's guidance [5].

HAMSTER: using Search Click Logs for Schema and Taxonomy Matching

In this Literature, we analyse an unsupervised matching of schema information from a large number of data sources into the schema of a data warehouse. The matching process is the first step of a framework to integrate data feeds from third- party data providers into a structured-search engine's data warehouse. We utilize technique based on the search engine's click logs.

Two schema elements are matched if the distributions of keyword queries that cause click-through on their instances are similar [6].

3. Proposed Model

In this section, we describe the automated knowledge base integration mechanism using principle component analysis technique on class structure integration and instance matching. This process is represented as follows

Representation of the Knowledge Base

A knowledge base (KB) is a tuple denoted by (E; L;R; P), consisting of a collection of entities E, literals L, relations R holding between entities and properties P holding between entities and literals. An entity e € E can be a class or an instance. E ={c U I, where C and I represent a class set and an instance set.

The example of KB, there is four entities - two classes, "Actor" and "Celebrity" and two instances, "Vijay" and "Sangeetha"; the date "24-6-1974" and string "Joseph Vijay" are literals; three relations "subclass", "type of" and "married to" and two properties "born" and "full name". The figure 1 describes the representation of the KB.

In a knowledge base, the classes form a hierarchical structure, in which different classes have "subclass/ superclass" relationships.

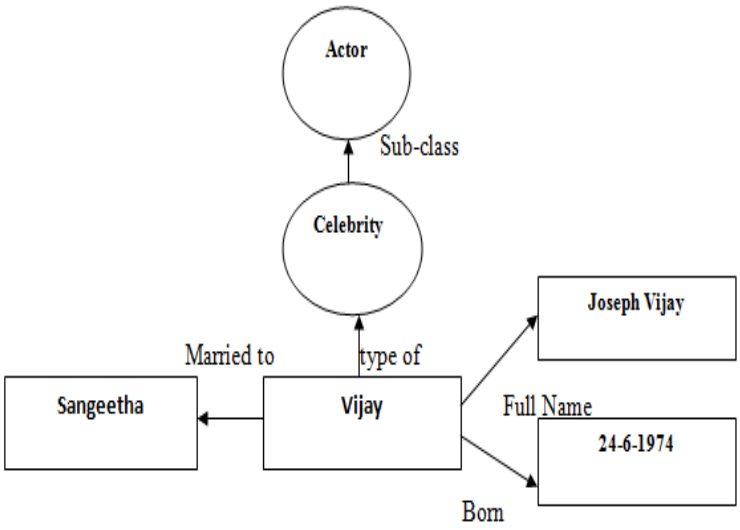

Figure 1: Representation of KB

In order to integrate two KBs, the projection of the work is to identify the positions of entities from one KB in another one to construct a unified KB. Other words Knowledge base integration is the process of identifying the position of each entity from Kb1 in Kb2 (or vice versa) to get the unified knowledge base.

3.2. Taxonomy or Class Integration

Taxonomy integration is the process of identifying the position of each class node from T1 in T2 to get the unified taxonomy.

We introduce class semantic relationship between classes into categories

Equivalence - The equivalence between classes refers that the two classes represent the same concept.

Generalization & Specification -The concept of one class is a subclass/super class of another one.

For Alignment of class, semantic relationship is has to be considered. Partitioning technique utilizes the ensemble mechanism to partition the entities, for an entity from KB1, if an equivalent relationship is found in KB2, then the position is identified.

Automatic Ensemble Framework for Integration of Ontology based Knowledge Bases

The proposed framework provides the Solution to the NP hard problem in terms of query selection.

1. Query Selection

Objective of taxonomy integration is to design a proper interface between the query and web data. Given two taxonomies T1 and T2, for each node in T1 we treat it as a query node and the position search space consists of all nodes of T2 (each node in T2 is called the target node). The contextual information of the target node has to be mapped to the query node.

The pruning strategy is applied according to the outcome of past queries, which may further influence the future query selection. Query selection strategy is to adaptively make a sequence of decisions. Adaptive query selection algorithm using greedy algorithms has been proposed to handle instance matching.

Algorithm 1: Query Selection
Input: Two Taxonomy T1 and T2
Output: Query Set Q
Process
Initialize Instance pair IP
Where Cp →0
Resultant Query Set
For each Instance Pair IP1 € IP
Generate query Q
Where Q= {LT1, i}
Q = arg Max (Q) € IP
Return Q

The Algorithm is adaptive greedy algorithm for query selection. The query set has associated with prior probability.

2. Class Structure Generation based on Contextual information

The conceptual information of the each node is been derived in the two knowledge bases. It considers the semantic heterogeneous class structures. It is the process of detecting pairs of matching entities among two large, clean but overlapping collections of entities.

Naive pairwise based instance matching is intractable for matching entities between two large KBs. In order to scale to large volumes of data, approximate techniques are adopted. Ensemble techniques cluster the similar entities into partitions.

The each class c from a taxonomy T1, first reduce the instance list to size of c by computing a prior belief of generalization/specification relationships and filter the classes with prior score lower than a threshold.

The detail Architecture of the proposed model is described in the figure 2.

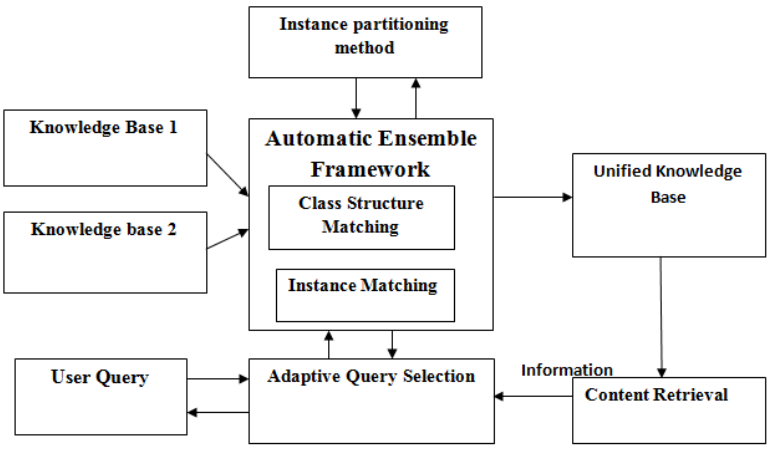

Figure 2: Architecture Diagram of the proposed Model

Ensemble framework produces the multiple class structures to the knowledge base as knowledge base is large in size and structure matching model is leverages to identify the relationship based on semantic in order to integrate the complex structures of the different KBs.

After normalizing the prior belief, we can get a probability distribution of results. Note that in our dataset, the KBs use the OWL identifier [7] to represent the instances, it is easy to get the instance equivalence information (this knowledge will be hidden when matching instances); for other dataset, the equivalence can be estimated using label information of instances as adopted. Initial instance matching pairs by considering the instance string representation. The similarity of a pair of candidate entities is computed using lexical similarities between entity names.

Ontology Integration on Semantic Knowledge Base

Ontology is an explicit specification of the conceptualization of a domain. Information models (such as the HL7 RIM) and standardized vocabularies (such as UMLS) can be part of ontology. Ontology provides a core component in a Knowledge-Based System. Ontology Integration can also carried out using Meta data[8]. Metadata is the detailed description of the instance data; the format and characteristics of the populated instance data; instances and values dependent on the requirements/role of the metadata recipient[8].

Once ontology Tags are obtained for the semantic embedded information in OWL file, the system will need to compare and merge this instance to gather more domain representation for the concepts in Semantic Knowledge base. Ontology integration [10] is to bridge conceptual model which represented lexical word on overlapping instance of the knowledge Base. Knowledge base integrated using principle component analysis[11].

OWL is a language for defining Web Ontologies [12] and their associated Knowledge Bases. The Knowledge integration using ontology is associated with the discriminant between the sub class can be achieved easily. The example is represented below

There are two types of animals, Male and Female.

```
<rdfs: Class rdf: ID="Male">
  <rdfs:subClassOf rdf:resource="#Animal"/>
</rdfs:Class>
```

The subClassOf element asserts that its subject - Male - is a subclass of its object -- the resource identified by #Animal.

```
<rdfs:Class rdf:ID="Female">
  <rdfs: subClassOf rdf: resource="#Animal"/>
  <owl: disjointWith rdf: resource="#Male"/>
</rdfs:Class>
```

One animal are Female, too, but nothing can be both Male and Female (in this ontology) because these two classes are disjoint (using the disjoint With tag).

4. Experimental Results

In section, we describe the experimental results of the proposed framework against the existing approaches. The experimental result demonstrates that proposed model outperforms the state of art approaches in terms of effectiveness, efficiency and accuracy. The detailed description is as follows

Dataset Description

We have done extensive experiments on 2 real datasets which is as follows

1.YAGO

YAGO is a Semantic knowledge base, in which entities, facts, and events are anchored in both time and space. YAGO2 is built automatically from Wikipedia, GeoNames, and Wordnet. It contains 447 million facts about 9.8 million entities. Human evaluation confirmed an accuracy of 95% of the facts in YAGO [9].

2. DBpedia

DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against datasets derived from Wikipedia and to link other datasets on the Web to Wikipedia data. We describe the extraction of the DBpedia datasets, and how the resulting information is published on the Web for human-and machine-consumption. We describe some emerging applications from the DBpedia community and show how website authors can facilitate DBpedia content within their sites.

Evaluation

The proposed Framework is evaluated against the following measures against several preprocessing steps on those data sets

1.Precision

Positive predictive value is the fraction of relevant instances among the retrieved instances. Precision is the number of correct feature divided by the number of all returned feature space.

$$\text{Precision} = \frac{\text{True positive}}{True\ positive + False\ Positive}$$

True positive is a number of real positive cases in the data and false negative is number of real negative cases in the data.

The precision is evaluated against different dataset is depicted in the figure 3 and performance values is described in the table 1 for all the dataset used in this work

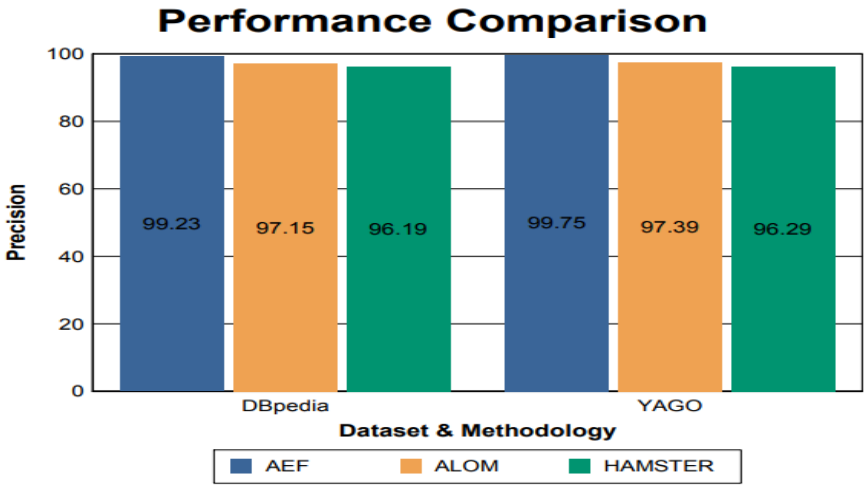

Figure 3: Performance Evaluation of the Methodologies on Precision against the Different Datasets

2. Recall

It is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. The recall is the part of the relevant documents that are successfully classified into the exact classes

$$\text{Recall} = \frac{\text{True positive}}{True\ positive + False\ negative}$$

True positive is a number of real positive cases in the data and false negative is number of real negative cases in the data. The recall is evaluated against different dataset is depicted in the figure 4 and performance values is described in the table 1 for all the dataset used in this work

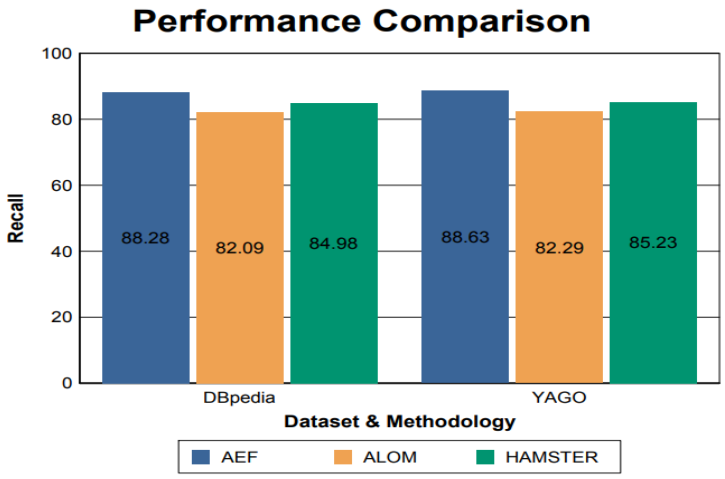

Figure 4: Performance Evaluation of the Methodologies on Recall against the Different Datasets

3. F Measure

It is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test. The Performance of the methodology is described in the figure 5 and performance values is described in the table 1 for all the dataset used in this work

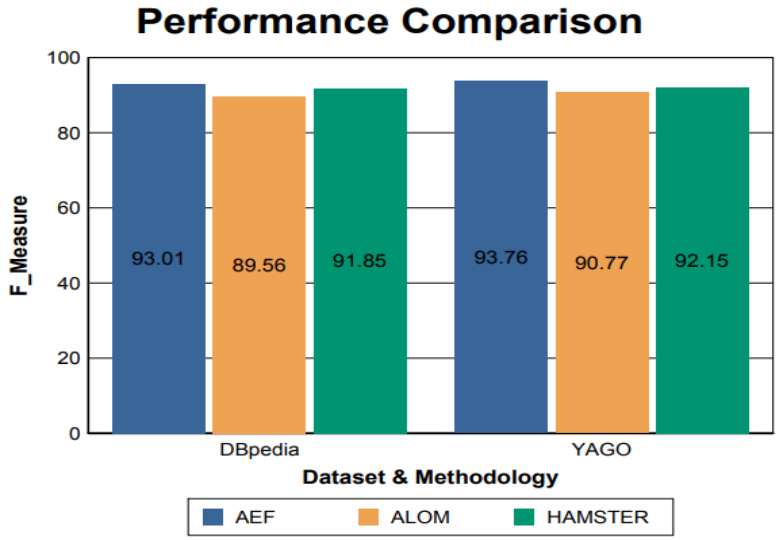

Figure 5: Performance Evaluation of the Methodologies on F Measure against the Different Datasets

4. Computation Time

It is defined as no of time taken to establish the instance matching for the different lexical words between the two heterogeneous sources.

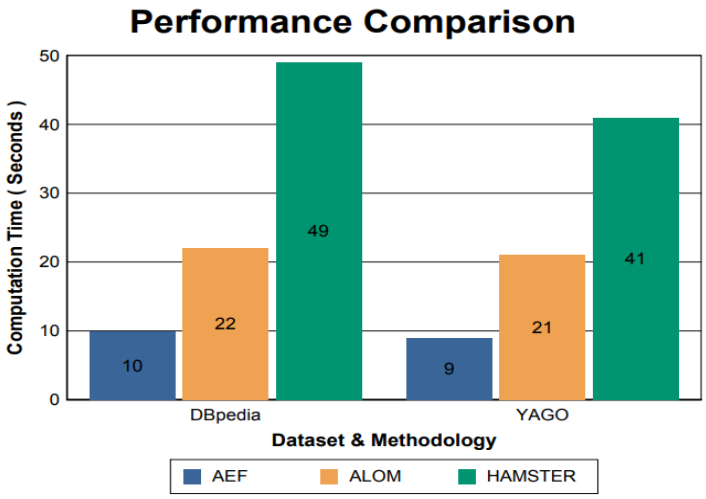

Figure 6: Performance Evaluation of the Methodologies on Computation Time against the Different Datasets

The performance evaluation chart of the computation time and its values is described in figure 6 and Table 1.

$$\text{Computation Time} = \frac{\text{No of time taken for Single instance}}{\text{Total time taken for entire Instance Mapping}}$$

Table 1: Performance Comparison of Methodology against Measures for Various Dataset

Dataset	**System**	**Precision in %**	**Recall In %**	**F Measure in %**	**Computation Time in seconds**
YAGO	AEF	99.75	88.63	93.76	9
	ALOM	97.39	82.29	90.77	21
	HAMSTER	96.29	85.23	92.15	41
DBpedia	AEF	99.23	88.28	93.01	10
	ALOM	97.15	82.09	89.56	22
	HAMSTER	96.19	84.98	91.85	49

The evaluation of result is described in the table 1 for DBpedia and YAGO datasets. It is observed that the proposed method is always better when compared to Class structure integration and with entity mapping using ontology tags; it has provided better or comparable results.

5. Conclusion

We have designed and implemented an Automatic Ensemble framework for Integration of Ontology Based Sematic Knowledge Base. The Problem of Knowledge base integration has been achieved with high accuracy. The greedy

based algorithm is also modelled to for query pruning. Based on the taxonomy integration result, we align the instance through an Ensemble constrainst and OWL constrainst. It has capability integrated the complex structures of the representation. Finally proposed system is verified to working better through extensive results in terms of both accuracy and efficiency.

References

[1] Hoffart J., Suchanek F.M., Berberich K., Weikum G., YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia, Artificial Intelligence 194(2) (2013), 8-61.

[2] Wu W., Li H., Wang H., Zhu K.Q., Pro-base: A probabilistic taxonomy for text understanding, In Proceedings of the ACM SIGMOD International Conference on Management of Data (2012), 481-492.

[3] Suchanek F.M., Abiteboul S., Senellart P., Paris: Probabilistic alignment of relations, instances, and schema, Proceedings of the VLDB Endowment 5(3) (2011), 157-168.

[4] Lacoste-Julien S., Palla K., Davies A., Kasneci G., Graepel T., Ghahramani Z., Sigma: Simple greedy matching for aligning large knowledge bases, In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (2013), 572-580.

[5] Shi F., Li J., Tang J., Xie G., Li H., Actively learning ontology matching via user interaction, The Semantic Web-ISWC (2009), 585-600.

[6] Giaretta P., Guarino N., Ontologies and knowledge bases towards a terminological clarification, Towards very large knowledge bases: knowledge building & knowledge sharing 25 (1995).

[7] Jean-Mary Y.R., Shironoshita E.P., Kabuka M.R., Ontology matching with semantic verification, Web Semantics: Science, Services and Agents on the World Wide Web 7(3) (2009), 235-251.

[8] Nandi A., Bernstein P.A., HAMSTER: using search click logs for schema and taxonomy matching, Proceedings of the VLDB Endowment 2(1) (2009), 181-192.

[9] Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., Ives Z., Dbpedia: A nucleus for a web of open data, The semantic web (2007), 722-735.

[10] Papadakis G., Ioannou E., Niederée C., Fankhauser P., Efficient entity resolution for large heterogeneous information spaces, In Proceedings of the fourth ACM international conference on Web search and data mining (2011), 535-544.

[11] Kondreddi S.K., Triantafillou P., Weikum G., Combining information extraction and human computing for crowd sourced knowledge acquisition, IEEE 30th International Conference on Data Engineering (ICDE) (2014), 988-999.

[12] Shvaiko P., Euzenat J., Ontology matching: state of the art and future challenges, IEEE Transactions on knowledge and data engineering 25(1) (2013), 158-176.