Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 12 Issue 10 Version 1.0 May 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Knowledgebase Representation for Royal Bengal Tiger in the Context of Bangladesh

By Md.Sarwar Kamal & Sonia Farhana Nimmy
BGC Trust University Bangladesh, Chittagong

Abstract - Royal Bengal Tiger is one of the penetrating threaten animal in Bangladesh forest at Sundarbans. In this work we have had concentrate to establish a robust Knowledgebase for Royal Bengal Tiger. We improve our previous work to achieve efficiency on knowledgebase representation. We have categorized the tigers from others animal from collected data by using Support Vector Machines(SVM) .Manipulating our collected data in a structured way by XML parsing on JAVA platform. Our proposed system generates n-triple by considering parsed data. We proceed on an ontology is constructed by Protégé which containing information about names, places, awards. A straightforward approach of this work to make the knowledgebase representation of Royal Bengal Tiger more reliable on the web. Our experiments show the effectiveness of knowledgebase construction. Complete knowledgebase construction of Royal Bengal Tigers how the efficient out-put. The complete knowledgebase construction helps to integrate the raw data in a structured way. The outcome of our proposed system contains the complete knowledgebase. Our experimental results show the strength of our system by retrieving information from ontology in reliable way.

Keywords : Ontology, Linked data, Web Semantics, XML parsing, N-triples, Royal Bengal Tiger.

GJCST-C Classification: I.2.4

KNOWLEDGEBASE REPRESENTATION FOR ROYAL BENGAL TIGER IN THE CONTEXT OF BANGLADESH

Strictly as per the compliance and regulations of:

Knowledgebase Representation for Royal Bengal Tiger in the Context of Bangladesh

Md.Sarwar Kamal [α] & Sonia Farhana Nimmy [σ]

Abstract - Royal Bengal Tiger is one of the penetrating threaten animal in Bangladesh forest at Sundarbans. In this work we have had concentrate to establish a robust Knowledgebase for Royal Bengal Tiger. We improve our previous work to achieve efficiency on knowledgebase representation. We have categorized the tigers from others animal from collected data by using Support Vector Machines(SVM) .Manipulating our collected data in a structured way by XML parsing on JAVA platform. Our proposed system generates n-triple by considering parsed data. We proceed on an ontology is constructed by Protégé which containing information about names, places, awards. A straightforward approach of this work to make the knowledgebase representation of Royal Bengal Tiger more reliable on the web. Our experiments show the effectiveness of knowledgebase construction. Complete knowledgebase construction of Royal Bengal Tigers how the efficient out-put. The complete knowledgebase construction helps to integrate the raw data in a structured way. The outcome of our proposed system contains the complete knowledgebase. Our experimental results show the strength of our system by retrieving information from ontology in reliable way.

IndexTerms : Ontology, Linked data, Web Semantics, XML parsing, N-triples, Royal Bengal Tiger.

I. Introduction

The sovereign Royal Bengal Tiger is drifting near the frontier of extinction. Once, the tiger cracked the whip over a supreme part of the globe ranging from the Pacific to the Black Sea and from Ural Mountains to the Mountain Agung. It is a paradox of fate that tiger is facing an assailment of poaching throughout its range. The main factor contributing in the decline of cat population is habitat degradation. But poaching has put them in a vulnerable condition to survive. The forest department sources said the big cat species are now disappearing fast from the world as the current population of tiger is only about 3700, down from around one lakh in 1900.There are only five sub-species of tigers surviving in the world which are Bengal tiger, Siberian tiger, Sumatran tiger, South- China tiger and Indo-China tiger. Balinese tigers, Javanese tigers and Caspian tigers have already vanished from the planet as the experts estimated that the remaining species of the big cat are likely to disappear immediately with the advent of next century. Official sources said at least 60 tigers were killed in the last three decades as the animals came to the nearby locality in search of food. According to review of the ministry, the big cats kill 25 to40 people annually while two to three tigers fall victim of mass-beating. According to a study conducted jointly by the United Nations, Bangladeshi government and Indian government in 2004, as many as 440 tigers have been found in the Bangladeshi part of the Sundarbans, the sources said. Right now tigers occupy only 7% of their historic range and they live in small islands of forests surrounded by a sea of human beings. Over the past few centuries tigers lost more than 80% of their natural habitats and what remain are only small fragments under heavy anthropogenic pressure.

This paper Organized as follows. In section II we have narrates Knowledgebase and Ontological basics and terminology which are essential for representation of Knowledgebase. In section III we described the General terminologies of Knowledgebase. In section IV we have described briefly Support Vector Machines (SVM) on the eve of categorized the Tiger from other animals. In section V we have elaborate INTRINSIC INFORMATION CONTENT METRIC and in next section we cited the Instance Matching Algorithm. last but not the least we have rape out by defining the challenges of the Ontology Instances Matching.

II. Knowledgebase and Ontology

Knowledge bases are playing an increasingly important role in enhancing the intelligence of Web and enterprise search and in supporting information integration. Today, most knowledge bases cover only specific domains, are created by relatively small groups of knowledge engineers, and are very cost intensive to keep up-to-date as domains change. At the same time, Wikipedia has grown intoone of the central knowledge sources of mankind, maintained by thousands of contributors Kobilarovetal. Collected data are organized to parsing and enable them to extract easily on the web. The complete knowledgebase contain information about Royal Bengal Tiger to enrich it. This knowledgebase helps to get informative knowledge about Royal Bengal Tiger who are an important part of our country as well as whole world. Our motivation is to provide a perfect representation of Royal Bengal Tiger on the web through Knowledgebase. The knowledge captured in the ontology can be used to parse and generate N-triples.

Author α : Lecturer, Computer science and engineering, BGC Trust University Bangladesh, Chittagong.E-mail : sarwar.bgctub@gmail.com
Author σ : Lecturer, Computer science and engineering, BGC Trust University Bangladesh, Chittagong. E-mail : nimmy_cu@yahoo.com

Structured data is easy to extract on the web which can be accessible for people to reach their goal. Our motive is to take the data in a structured way.

a) Ontology Alignment

Alignment A is defined as a set of correspondences with quadruples < e; f; r; l > where **e** and **f** are the two aligned entities across ontology's, **r** represents the relation holding between them, and **l** represents the level of confidence [0, 1] if there exists in the alignment statement. The notion **r** is a simple (one-to-one equivalent) relation or a complex (subsumption or one-to-many) relation Ehrig (2007). The correspondence between **e** and **f** is called aligned pair throughout the paper. Alignment is obtained by measuring similarity values between pairs of entities.

The main contribution of our Anchor-Flood algorithm is of attaining performance enhancement by solving the scalability problem in aligning large ontology's. Moreover, we obtain the segmented alignment for the first time in ontology alignment field of research. We achieve the best runtime in world-wide competitions organized by Ontology Alignment Evaluation Initiative (OAEI) 2008 (held in Karlsruhe, Germany) and 2009 (held in Chantilly, VA, USA).

b) Intrinsic Information Content

We propose a modified metric for Intrinsic Information Content (IIC) that achieves better semantic similarity among concepts of ontology. The IIC metric is integrated with our Anchor-Flood algorithm to obtain better results efficiently.

c) Ontology and Knowledge Base

According to Ehrig (2007), an ontology contains core ontology, logical mappings, a knowledge base, and a lexicon. A core ontology, S, is defined as a tuple of five sets: concepts, concept hierarchy or taxonomy, properties, property hierarchy, and concept to property function.

$$S = (C, \leq c\ R, \sigma\ , \leq R)$$

where **C** and **R** are two disjoint sets called concepts" and relations" respectively. A relation is also known as a property of a concept. A function represented by σ(r) =< dom(r); ran(r) > where r € R, domain is dom(r) and range is ran(r). A partial order≤R represents on R, called relation hierarchy, where r1≤R r2 iff dom (r1) ≤C dom (r2) and ran (r1) ≤C ran (r2). The notation ≤C represents a partial order on C, called concept hierarchy or taxonomy". In a taxonomy, if c1 <C c2 for c1; c2€C, then c1 is a sub concept of c2, and c2 is a super concept of c1. If c1 <C c2 and there is no c3€C with c1 <C c3 <C c2, then c1 is a direct sub concept of c2, and c2 is a direct super concept of c1 denoted by c1 ≺ c2. The core ontology formalizes the intentional aspects of a domain. The extensional aspects are provided by knowledge bases, which contain asserts about instances of the concepts and relations. A knowledge base is a structure KB = (C,R, I, C, , R) consisting of

_ two disjoint sets C and R as defined before,

_ a set I whose elements are called instance identifiers (or instance for short),

_ a function C : C→ Ə(I) called concept instantiation,

_ a function { R: R → Ə(I2) with (r) ⊆ ᶅC (dom(r)) × ᶅC (ran(r)), for all r ɛ R. The function R is called relation instantiation.

With data types being concepts as stated for core ontology, concrete values are analogously treated as instances.

III. General Terminology

This section introduces some basic definitions of terminologies of semantic web to familiarize the readers with the notions used throughout the paper. It includes the definitions of ontology and knowledgebase, linked data, Geonames, Geospatial data, and N-triples from semantic web to comprehend the essence of our paper.

a) N-Triples

N-Triples is a format for storing and transmitting data. It is a line-based, plain text serialization format for RDF (Resource Description Framework) graphs, and a subset of the Turtle (Terse RDF Triple Language) format.[1][2] N-Triples should not be confused with Notation 3 which is a superset of Turtle. N-Triples was primarily developed by Dave Beckett at the University of Bristol and Art Barstow at the W3C. N-Triples was designed to be a simpler format than Notation 3 and Turtle, and therefore easier for software to parse and generate. However, because it lacks some of the shortcuts provided by other RDF serializations (such as CURIEs and nested resources, which are provided by both RDF/XML and Turtle) it can be onerous to type out large amounts of data by hand, and difficult to read.

b) Geonames

Geonames is a geographical database available and accessible through various Web services, under a Creative Commons attribution license. Geonames is integrating geographical data such as names of places in various languages, elevation, population and others from various sources. All lat/long coordinates are in WGS84 (World Geodetic System 1984). Users may manually edit, correct and add new names using a user friendly wiki interface.

c) Geospatial Data

Geospatial data is information that identifies the geographic location and characteristics of natural or constructed features and boundaries on the earth, typically represented by points, lines, polygons, and or complex geographic features. This includes original and interpreted geospatial data, such as those derived through remote sensing including, but not limited to, images and raster data sets, aerial photographs, and other forms of geospatial data or data sets in both digitized and non-digitized forms.

d) Neighbouring of Geospatial Data

At first, we find the neighbours of a division. In the same way we also find the neighbours of other six divisions. After that, we find the neighbours of all districts. At last, we find the neighbours of all sub districts one by one.

e) Linked Data

With the structures of ontology and ontology knowledge base, semantic web visionaries coined the term linked data, which uses Resource Description Framework (RDF) and RDF triples to connect related instances. The term refers to a style of publishing and interlinking structured data on the Web. The basic assumption behind Linked Data is that the value and usefulness of data increases the more it is interlinked with other data. In summary, Linked Data is simply about using the Web to create typed links between data from different sources. However, semantic knowledge base and linked data is used synonymously throughout this paper.

f) Semantic Web

The Semantic Web1 has received much attention recently. Its vision promises an extension of the current web in which all data is accompanied with machine understandable metadata allowing capabilities for a much higher degree of automation and more intelligent applications (Berners-Lee et al., 2001). To make this idea more concrete, consider the statement The University of Georgia is located in Athens, GA. To a human with knowledge of colleges and universities and the geography of the southeastern United States, the meaning of this statement is clear. In addition, upon seeing this statement, other related information comes to mind such as professors who work at the University. In a Semantic Geospatial Web context (Egenhofer, 2002), this related information would be GIS data and services, such as road network data and facility locations for the Athens area which could be combined with way finding services. The goal of the Semantic Web is to make the semantics of such data on the web equally clear to computer programs and also to exploit available background knowledge of related information. On the Semantic Web this statement would be accompanied with semantic metadata identifying an instance of the concept University with the name The University of Georgia. Similarly, the instance of City and State, Athens, GA, would unambiguously describe the university's geographic location. Note the distinction between semantic metadata describing high-level concepts and relationships and syntactic and structural metadata describing low level properties like file size and format. To create this semantic metadata, we must identify and mark occurrences of known entities and relationships in data sources. This tagging process is known as metadata extraction and semantic annotation. These annotations are especially important for multimedia data, as non textual data has a very opaque relationship with computers. Some examples of annotation of textual and multimedia data are presented in (Dill et al., 2003; Hammond et al. 2002), and (Jin et al., 2005) respectively. To provide ontological metadata in a machine process able form, a standard way to encode it is needed. The W3C has adopted Resource Description Framework (RDF) as the standard for representing semantic metadata. Metadata in RDF is encoded as statements about resources. A resource is anything that is identify able by a Uniform Resource Identifier (URI). Resources can be documents available on the web or entities which are not web-based, such as people and organizations.

IV. Support Vector Machines

Support Vector Machine (SVM) is one of the latest clustering techniques which enables machine learning concepts to amplify predictive accuracy in the case of axiomatically diverting data those are not fit properly. It uses inference space of linear functions in a high amplitude feature space, trained with a learning algorithm. It works by finding a hyperplane that linearly separates the training points, in a way such that each resulting subspace contains only points which are very similar. First and foremost idea behind Support Vector Machines (SVMs) is that it constituted by set of similar supervised learning. An unknown tuple is labeled with the group of the points that fall in the same subspace as the tuple. Earlier SVM was used for Natural Image processing System (NIPS) but now it becomes very popular is an active part of the machine learning research around the world. It is also being used for pattern classification and regression based applications. The foundations of Support Vector Machines (SVM) have been developed by V.Vapnik.

Two key elements in the implementation of SVM are the techniques of mathematical programming and kernel functions. The parameters are found by solving a quadratic programming problem with linear equality and inequality constraints; rather than by solving a non-convex, unconstrained optimization problem. The flexibility of kernel functions allows the SVM to search a wide variety of hypothesis spaces. All hypothesis space help to identify the Maximum Margin Hyperplane(MMH) which enables to classify the best and almost correct data The following figure shows the process of SVMs selection from large amount of SVMs.

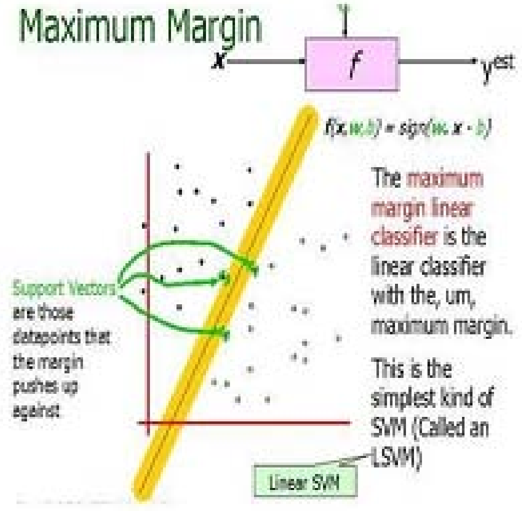
Maximum Margin
x
f
y
y^est
f(x,w,b) = sign(w. x - b)
The maximum margin linear classifier is the linear classifier with the, um, maximum margin.
This is the simplest kind of SVM (Called an LSVM)
Support Vectors are those datapoints that the margin pushes up against
Linear SVM

Fig 1: Maximum Margin Hyper Plane

Expression for Maximum margin is given as [4][8] (for more information visit [4]

$$\text{margin} \equiv \underset{\mathbf{x}\in D}{\arg\min}\, d(\mathbf{x}) = \underset{\mathbf{x}\in D}{\arg\min} \frac{|\mathbf{x}\cdot\mathbf{w}+b|}{\sqrt{\sum_{i=1}^{d} w_i^2}}$$

The above illustration is the maximum linear classifier with the maximum range. In this context it is an example of a simple linear SVM classifier. Another interesting question is why maximum margin? There are some good explanations which include better empirical performance. Another reason is that even if we've made a small error in the location of the boundary this gives us least chance of causing a misclassification. The other advantage would be avoiding local minima and better classification. Now we try to express the SVM mathematically and for this tutorial we try to present a linear SVM. The goals of SVM are separating the data with hyper plane and extend this to non-linear boundaries using kernel trick [8] [11]. For calculating the SVM we see that the goal is to correctly classify all the data. For mathematical calculations we have,

[a] If Yi= +1;

[b] If Yi= -1; wxi + b ≤ 1

[c] For all i; yi (wi + b) ≥ 1

In this equation x is a vector point and w is weight and is also a vector. So to separate the data [a] should always be greater than zero. Among all possible hyper planes, SVM selects the one where the distance of hyper plane is as large as possible. If the training data is good and every test vector is located in radiusr from training vector. Now if the chosen hyper plane is located at the farthest possible from the data [12]. This desired hyper plane which maximizes the margin also bisects the lines between closest points on convex hull of the two datasets. Thus we have [a], [b] & [c].

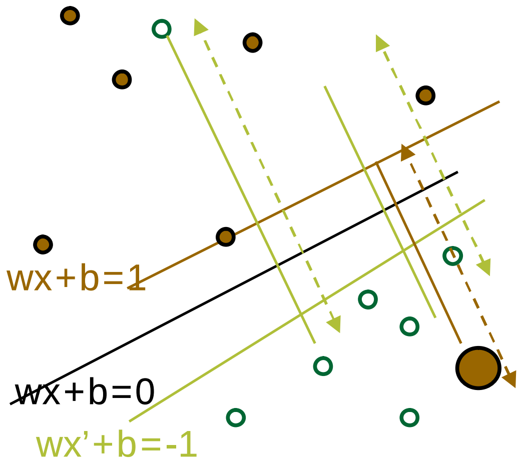

Figure 2 : Representation of Hyper planes

Distance of closest point on hyperplane to origin can be found by maximizing the x as x is on the hyper plane. Similarly for the other side points we have a similar scenario. Thus solving and subtracting the two distances we get the summed distance from the separating hyperplane to nearest points. Maximum Margin = M = 2 / ||w||

V. Related Work

Before this work we have had work to prepare ontology for medical document classification. We have reviewed 20 research journals on the eve of knowledgebase representation for Tigers but we got only a few that does not indicates the outcome for Tigers knowlegebase.

VI. Proposed Modification In Intrinsic Information Content Metric

To overcome the limitation of the state-of-art metrics of computing semantic similarity among concepts within domain ontology and to cope with the new ontologies with the introduced complex description logics, we propose a modified metric of computing intrinsic information content. The metric can be applied to a simple taxonomy and to a recent complex OWL ontology as well.

The primary source of IC in ontology is obviously concepts and concept hierarchy. However, OWL ontology also contains properties, restrictions and other logical assertions, often called as relations. Properties are used to define functionality of a concept explicitly to specify a meaning. They are related to concept by means of domain, range and restrictions.

According to Resnik, semantic similarity depends on the shared information. As Resnik introduces the IC which represents the expressiveness of a particular concept. Classical metric of IC are based on the available concepts in taxonomy or in a large text corpora. However, as time passes on, the definition and the content of ontology becomes more and more complex. The expressiveness of a concept is not only rely on the concept taxonomy but also on the other relations like properties and property-restrictions.

We already have discussed about the probable sources of information content(IC) or the expressiveness of semantic similarity among the concepts of ontology. We find that the IC of a concept is negatively related to the probability of a concept in external large text corpora Resnik (1995). We also find that the IC of a concept is inversely related to the number of hyponyms or the concepts it subsumes Seco et al. (2004). Moreover, we observe that description logic (DL) based ontology of semantic technology is formal and explicit in its conceptualization with the help of relations. Every concept is defined with sufficient semantic embedding with the organization, property functions, property restrictions and other logical assertions. Current ontology of semantic technology is defined as an explicit specification of a conceptualization" Gruber (1995). Although the most domain ontologies are not as complete as Word Net in terms of concepts and concept organization, they have well support from logical assertions to define a concept concisely. Therefore, we can obtain sufficient IC of a concept without depending on the external large text corpora heavily, required that we use intrinsic information of the concept. One of the good sources of intrinsic information of a concept is its relations by means of property functions and property restrictions. Our relation based IC is defined as:

Icrel(c)

$$\mathrm{Ic}_{\mathrm{rel}}(\mathrm{c}) = \frac{Log(rel(C)+1)}{Log(total\_rel+1)} \qquad (1)$$

Where rel stands for the relation of properties, property function and restrictions, rel(c) denotes the number of relations of a concept c and total rel represents the total number of relations available in the ontology.

As long as the information content of a concept depends both on the hyponyms or sub sumption relations of a concept and the related properties of the concept, we need to integrate the icre(c) with the Seco's metric This integration introduces a coefficient factor ρ and the equation becomes as:

$$ic(c) = \rho.icrel(c) + (1-\rho).\ icseco(c) \qquad (2)$$

Concepts	Number of relations	Number of Hypotenuse	IC $_{scco}$	IC $_{rel}$	IC $_{modified}$
Date	3	0	1.000	0.332	0.641
Page Range	2	0	1.000	0.263	0.603
Organization	0	3	0.613	0.000	0.283
Institution	3	1	0.693	0.322	0.257
Publisher	3	2	1.000	0.222	0.123
School	3	3	1.000	0.123	0.968
List	0	0	1.000	0.258	0.987
Person List	4	0	1.000	0.125	0.789
Journal	2	1	1.000	0.236	0.456
Address	3	0	1.000	0.125	0.489
Person	0	1	1.000	0.000	0.478
Conference	0	3	1.000	0.231	0.258
Reference	1	0	1.000	0.963	0.369
Academic	6	0	1.000	0.000	0.123
PhDThesis	5	1	1.000	0.217	0.147
MastersThesis	2	2	1.000	0.235	0.258
Misc	0	2	0.873	0.148	0.000
Motion	0	2	0.521	0.148	0.123
Picture Part	0	3	0.123	0.000	0.236
In Collection	0	0	1.000	0.789	0.214

Table 1: contains IC values measured by Saco's metric and our modified metric

Where the coefficient factor ρ is defined by the nature of ontology. While a small size of ontology is often incomplete by its concepts alone, the coefficient factor tends to increase to focus on relations. On the contrary, when relations are inadequate to define a concept and there are a large number of concepts in the taxonomy, ρ tends to decrease its value. However, we definitely need a trade-off to select the coefficient factor and we define it as:

$$\rho = \frac{Log(total\_rel+1)}{Log(total\_rel)+Log(total\_concept)}$$

Where total_rel is the maximum number of relations while total_ concepts is the maximum number of concepts available in an ontology.

From the experiments, we also observe that the deeper concepts have more expressiveness or larger IC values. Therefore, it guarantees that our modified IC metric takes the depth of a concept implicitly and the children of a concept explicitly.

However, we do not take the link type and local concept density into account unlike expressed in Jiang & Conrath (1997). As we consider thyponyms by incorporating the Saco's IC metric, it considers the edges between sub sumption concepts implicitly Furthermore, we also compute semantic similarity for every possible pair of concepts of the ontology.

e1	e2	Sim_{seco}	$Sim_{proposed}$
Reference	PhD Thesis	0.113	0.782
Reference	Master's Thesis	0.113	0.782
Reference	In Collection	0.113	0.782
Reference	In Proceedings	0.113	0.782
Reference	Article	0.113	0.790
Reference	Chapter	0.113	0.784
Reference	In Book	0.113	0.784
Reference	TechReport	0.113	0.777
Reference	Deliverable	0.113	0.784
Reference	Manual	0.113	0.790
Reference	Unpublished	0.113	0.790
Reference	Booklet	0.113	0.777
Reference	Lecture Notes	0.113	0.788
Reference	Collection	0.113	0.782
Reference	Monograph	0.113	0.782
Reference	Proceeding	0.113	0.782

Table 2: contains semantic similarity between Reference to each of its leaves considering Saco's metric and our proposed metric

VII. Instance Matching Algorithm

The operational block of the instance matching integrates ontology alignment, retrieves semantic link clouds of an instance in ontology and measures the terminological and structural similarities to produce matched instance pairs. Pseudo code of the Instance Matching algorithm:

```
Algo. InstanceMatch (ABox ab1, ABox ab2, Alignment A)
for each insi element of ab1
    cloudi=makeCloud(insi,ab1)
    for each insj element of ab2
        cloudj=makeCloud(insj,ab2)
        if $\forall$a(c1; c2) elements of A|c1 elements of Block(ins1:type) ^
        c2 elements of Block(ins2:type)
            if Simstruct(cloudi; cloudj) $\geq \delta$
                imatch=imatch $\bigcup$ makeAlign(insi; insj)
```

VIII. Ontology Instance Matching Challenges

The ontology schema, which includes concepts, properties and other relations, is relatively stable part of an ontology. However, concepts and properties of ontology are instantiated very often by deferent users in deferent styles. Thus, ontology instances are dynamic in nature and are challenging to be matched. Structural variants compose of the most challenging variations in defining instances. To define an instance of a concept, ontology users usually take support from the properties, either object properties or data properties. Properties always behave like functions having domains and ranges. There might be a great variation of using property functions in their range values. The range of an Object Property is an instance while the range of a Data type Property is an absolute value. There is always a chance of defining an Object Property of ontology as a Data type Property in ontology and vice versa. The cases of defining aproperty by another instance in one ABox and defining the property by a value in other ABox yield a great challenge in instance matching.

a) Approach to Solve the Challenges

We resolve typographical variation by the methods of data cleansing. The task of data cleansing comprises the detection and resolution of errors and inconsistencies from a data collection. Typical tasks are syntax check, normalization, and error correction. First of all, our syntax check and normalization process check the data type of an instance and classify on three important information types: time data (using regular expression), location data (using Geo Names Web service) and personal data. In our current realization, we use a couple of manually defined normalization rules for each information type. We implemented the module in a modular way, so that the used algorithm and rules of normalization can be extended and substituted. In instance matching, we need to look up the type (concept as a type of an instance) match of instances first. To cope with the logical variation, we first look up a block of concepts that includes the original type of an instance against another block of concepts which includes the type of another instance to be compared with instead of comparing two types alone. A relational block is defined as follows:

Definition 1: As concepts are organized in a hierarchical structure called a taxonomy, we consider a relational block of a concept c as a set of concepts and simply referred to block throughout this paper, and defined as:

$$\text{block}(c) = \{\text{children}(c) \bigcup \text{siblings}(c) \bigcup \text{parents}(c) \bigcup \text{grandparents}(c)g\}$$

where children(c) and parents(c) represent the children and the parents of a particular concept c, respectively within a taxonomy, whereas siblings(c) is defined as children (parents(c)-c and grandparents(c) is defined as parents (parents(c)) In an ontology, neither a concept nor an instance comprises its full specification in its name or URI (Uniform Resource Identifier) alone. Therefore we consider the other semantically linked information that includes other concepts, properties and their values and other instances as well. They all together make an information cloud to specify the meaning of that particular instance. The degree of certainty is proportional to the number of semantic links associated to a particular instance by means of property values and other instances. We refer the collective information of association as a Semantic Link cloud (SLC), which is defined as below:

Definition 2: A Semantic Link Cloud (SLC) of an instance is defined as a part of knowledge base Ehrig

(2007) that includes all linked concepts, properties and their instantiations which are related to specify the instance sufficiently.

IX. Conclusions

In this dissertation, we described the Anchor-Flood algorithm that can align ontologies of arbitrary size effectively, and that makes it possible to achieve high performance and scalability over previous alignment algorithms. To achieve these goals, the algorithm took advantage of the notion of segmentation and allowed segmented output of aligned ontologies. Specifically, owing to the segmentation, our algorithm concentrates on aligning only small sets of the entire ontology data iteratively, by considering\locality of reference". This brings us a by-product of collecting more alignments in general, since similar concepts are usually more densely populated in segments. Although we need some further refinement in segmentation, we have an advantage over traditional ontology alignment systems, in that the algorithm finds aligned pairs within the segments across ontologies and it has more usability in different discipline of specific modelling patterns. When the anchor represents correct aligned pair of concepts across ontologies, our Anchor-Flood algorithm finds segmented alignment within conceptually closely connected segments across ontologies efficiently. Even if the input anchor is not correctly defined, our algorithm is also capable of handling the situation of re- porting misalignment error. The complexity analysis and a different set of experiments demonstrate that our proposed algorithm outperforms in some aspect to other alignment systems. The size of ontologies does not affect theefficiency of Anchor-Flood algorithm. The average complexity of our algorithm is ON log(N), where N is the average number of concepts of ontologies.

References Références Referencias

1. Benjamin's, V Contreras, J., Corcho, O. & G_omez-P_erez, A. (2004).Six Challenges for the Semantic Web. AIS SIGSEMIS Bulletin,
2. Berners-Lee, T., Fischetti, M. & Dertouzos, M. (1999). Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web. Harper San Francisco.
3. Caracciolo, C., Euzenat, J., Hollink, L., Ichise, R., Isaac, A.,Malais_e, V., Meilicke, C., Pane, J., Shvaiko, P., Stuckenschmidt, H., _Sv_ab-Zamazal, O. & Sv_atek, V. (2008). Results of the Ontology Alignment Evaluation Initiative 2008. Proceedings of Ontology Matching Work-shop of the 7th International Semantic Web Conference, Karlsruhe, Germany. Collins, A. & Quillian, M. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior.
4. Burges C., "A tutorial on support vector machines for pattern recognition", In "Data Mining and Knowledge Discovery". Kluwer Academic Publishers, Boston, 1998, (Volume 2).
5. Ehrig, M. (2007). Ontology Alignment: Bridging the Semantic Gap. Springer, New York.
6. Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. The MIT Press, Cambridge
7. Euzenat, J., Ferrara, A., Hollink, L., Joslyn, C., Malais_e, V.,Meilicke, C., Nikolov, A., Pane, J., Scharffe, F., Shvaiko, P.,Spiliopoulos, V., Stuckenschmidt, H., _Sv_ab-Zamazal, O., Sv_atek, V., Santos, C.T. & Vouros, G. (2009). Preliminary results of the Ontology Alignment Evaluation Initiative 2009. Proceedings of Ontology MatchingWorkshop of the 8th International Semantic Web Conference, Chantilly, VA,USA.
8. Nello Cristianini and John Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000
9. Image found on the web search for learning and generalization in svm following links given in the book above. Engineering Review.
10. Grau, B., Parsia, B., Sirin, E. & Kalyanpur, A. (2005). Modularizing OWL ontologies. Proc. KCAP-2005 Workshop on Ontology Management, Banff , Canada.
11. Tom Mitchell, Machine Learning, McGraw-Hill Computer science series, 1997.
12. J.P.Lewis, Tutorial on SVM, CGIT Lab, USC, 2004.
13. Hirst, G. & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. WordNet: An electronic lexical database.
14. Hu, W., Cheng, G., Zheng, D., Zhong, X. & Qu, Y. (2006a). The Results of Falcon-AO in the OAEI 2006 Campaign. Proceedings of Ontology Matching (OM-2006), Athens, Georgia, USA.
15. Hu, W., Zhao, Y. & Qu, Y. (2006b). Partition-based Block Matching of Large Class Hierarchies. Proceedings of the 1st Asian Semantic Web Conference (ASWC2006), Beijing, China.
16. Hu, W., Qu, Y. & Cheng, G. (2008). Matching Large Ontologies: A Divideand- Conquer Approach. Data and Knowledge Engineering
17. Jiang, J. & Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings on International Conference on Research in Computational Linguistics, Taiwan.
18. Knappe, R., Bulskov, H. & Andreasen, T. (2007). Perspectives on ontology-based querying. International Journal of Intelligent Systems.
19. Kobilarov, G., Bizer, C., Auer, S. & Lehmann, J. Dbpedia-a linked data hub and data source for web and enterprise applications. Programme Chairs.

20. Lin, D. (1998). An information-theoretic de_nition of similarity.
21. Melnik, S., Garcia-Molina, H. & Rahm, E. (2002). Similarity Flooding: AVersatile Graph Matching Algorithm and its Application to Schema Matching.Proceedings of the 18th International Conference on Data Engineering (ICDE2002), San Jose, CA, USA.
22. Miller, G. & Charles, W. (1991). Contextual Correlates of Semantic Similarity. Language and cognitive processes.
23. Mitschick, A., Nagel, R. & Mei_ner, K. (2008). Semantic Metadata Instantiation and Consolidation within an Ontologybased Multimedia Document Management System. In Proceedings of the 5th European Semantic Web Conference ESWC.
24. Rada, R., Mili, H., Bicknell, E. & Blettner, M. (1989). Development and application of a metric on semantic nets. IEEE transactions on systems, man and cybernetics.
25. Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada.
26. Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language.Journal of arti_cial intelligence.
27. Rogers, J. (2005). OpenGALEN: Making the Impossible Very Difficult.http://www.opengalen.org.
28. Rosse, C. & Mejino, J. (2003). A Reference Ontology for Biomedical Informatics: the Foundation Model of Anatomy. Journal of Biomedical Informatics.
29. Seco, N., Veale, T. & Hayes, J. (2004). An intrinsic information contentmetric for semantic similarity in WordNet. Proceedings of ECAIf2004, the 16th European Conference on Arti_cial Intelligence.
30. Seddiqui, M. & Aono, M. (2008). Alignment Results of Anchor-Flood Algorithm for OAEI-2008. Proceedings of Ontology Matching Workshop of the 7th International Semantic Web Conference, Karlsruhe, Germany.
31. Seddiqui, M.H. & Aono, M. (2009a). An E_cient and Scalable Algorithm for Segmented Alignment of Ontologies of Arbitrary Size. Journal of Web Semantics: Science, Services and Agents on the World Wide Web.
32. Seddiqui, M.H. & Aono, M. (2009b). Anchor-Flood: Results for OAEI-2009.Proceedings of Ontology Matching Workshop of the 8th International Semantic Web Conference, Chantilly, VA, USA.
33. Seidenberg, J. & Rector, A. (2006). Web Ontology Segmentation: Analysis, Classification and Use. Proceedings of the 15th International Conference on World Wide Web (WWW2006), Edinburgh, Scotland.
34. Shannon, C. & Weaver, W. (1948). A mathematical theory of communication. Bell System Technical Journal.
35. Shvaiko, P. & Euzenat, J. (2008). Ten challenges for ontology matching.Proceedings of the 7th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE), Monterrey, Mexico.
36. Stoilos, G., Stamou, G. & Kollias, S. (2005). A String Metric for Ontology Alignment. Proceedings of the 4th International Semantic Web Conference (ISWC2005), Galway, Ireland.
37. Stuckenschmidt, H. & Klein, M. (2004). Structure-based Partitioning of Large Concept Hierarchies.