Domain Ontology Construction by Partial Import for Document Annotation

Tayybah Kiren, Muhammad Shoaib, and Sang-Jo Lee

***Abstract*—Semantic annotation is a useful technique to understand the underlying meanings of the document. Domain specific knowledge provided by the domain ontologies is needed to semantically annotate the document. But the main problem is the availability of an appropriate ontology for the domain. Ontology construction from scratch is protracted and labor intensive job. Therefore, it is advantageous to construct the ontologies by reusing the existing ontologies. This paper proposes a technique for constructing domain ontology for semantic annotation of a document by partially importing the existing ontologies.**

***Index Terms*—Semantic annotation, partial reuse, domain ontology.**

I. Introduction

Finding the right information from the huge corpus of documents available on the World Wide Web is very difficult. Documents contain the valuable knowledge for particular domain. But this knowledge cannot be efficiently exploited by machines for automation purpose because of the unstructured nature of the documents. Thus the creation of semantic metadata related to document content seems to be a way to exploit this knowledge and extract implicit and explicit information. The process of creating semantic metadata is called "Semantic annotation".

Ontologies are defined as formal explicit specifications of shared conceptualizations. In Semantic web, domain ontologies are considered to be the major source of semantic metadata creation. But unfortunately constructing domain ontologies from scratch is very time consuming task which makes the semantic annotation process very slow. Hence there is a need of constructing the domain ontologies by partially reusing the existing ontologies on the web. This paper proposes an approach for constructing domain ontology for semantic annotation by partially reusing the existing relevant ontologies on the web. Section II presents the existing approaches for domain ontology construction and for semantic annotation, proposed approach is described in Section III and evaluated in Section IV. Section V presents the conclusion.

II. Related Work

Constructing domain ontologies from text document is a non-trivial task. Most of the existing approaches constructontology from scratch. Sleeman et al proposed an approach for extracting ontological knowledge from the existing knowledge bases [1]. In [2] text corpus is used as background knowledge for constructing ontologies. Problem with these approaches that the knowledge sources used for ontology building do not express the knowledge in explicit way. This can limit the constructed ontologies. Moreover the construction of ontologies from scratch is quite time consuming task. A solution to these problems is reusing the existing relevant ontologies. There are some approaches that construct the ontologies by importing the existing relevant ontologies.The approach presented in[3] partition the large ontologies into small portions according to their class hierarchy structure. Difficulty with this approach is that it is not suitable for small size ontologies. Grau *et al.* [4] introduced a module extraction approach based on logic. This approach extracts locality based modules. There are many approaches for ontology based semantic annotation of text documents. Most of these approaches use existing ontologies for semantic annotation. Ontologlass [5] is an ontology based annotation system that performs morpheme level annotation. Amaya [6] is performs manual semantic annotation using RDF data but it just provides information like author, creator or title of document. OntoAnnotate [7] is another annotation tool, which keeps a copy of document to be annotated along with the annotation in the annotation database, thus making the annotation database heavy. All these semantic annotation approaches use some existing ontology for making annotations. Thus it is evident from the literature that there are many approaches for ontology construction from scratch and ontology partitioning and modularization. But there is no complete methodology for constructing domain ontologies by partially reusing the existing ontologies and using them to semantically annotatingthe unstructured document.

III. Proposed Approach

Architecture of the proposed approach for domain ontology construction for semantic annotation is shown in Fig. 1 Preprocessing of document includes stemming and removing the common words as they are not good candidate to be the domain concepts. The domain concepts for which no relevant ontologies exist are defined specifically according to the domain ontology and others are retrieved in the form of relevant modules from existing ontologies. Stepwise description of the proposed methodology is given in the following text.

A. Step 1: Finding Domain Concepts

To construct the domain ontologies first we will find out the domain concepts from given document. Keywords are

Manuscript received July 9, 2013; revised September 16, 2013.

Tayybah Kiren is with University of Engineering and Technology Lahore, Pakistan (e-mail:Tayybah@uet.edu.pk).

born concepts in the domain to which they belong [8]. Hence we first find out the keywords from the input document. Keywords are extracted on the basis of format (whether it is bold or highlighted) and position in the document. Position of a word in the document also shows its significance and relatedness to the domain of document [9]. For example if a word appears both in introduction and conclusion it generally carries more information. Thus domain concepts are extracted by using Position weight formula as presented in equation 1 which extracts keyword based on linguistic features.

$$PW(t,d) = \sum_{i=1}^{n} pw(t_i) \quad (1)$$

where $pw(t_i)$ is calculated by taking into consideration the weights assigned to the paragraph

$$Rank(O) = \sum_{i=1}^{n} Match(O,k_i) \quad (2)$$

where $Match(O,k_i)$ =

1 if there is exact matchbetween O and K_i

0 if there is no match between O and K_i

0.5 if there is match between O and K_i throughWordNet synonym

and sentence to which the word *ti* belongs, and d is the document.

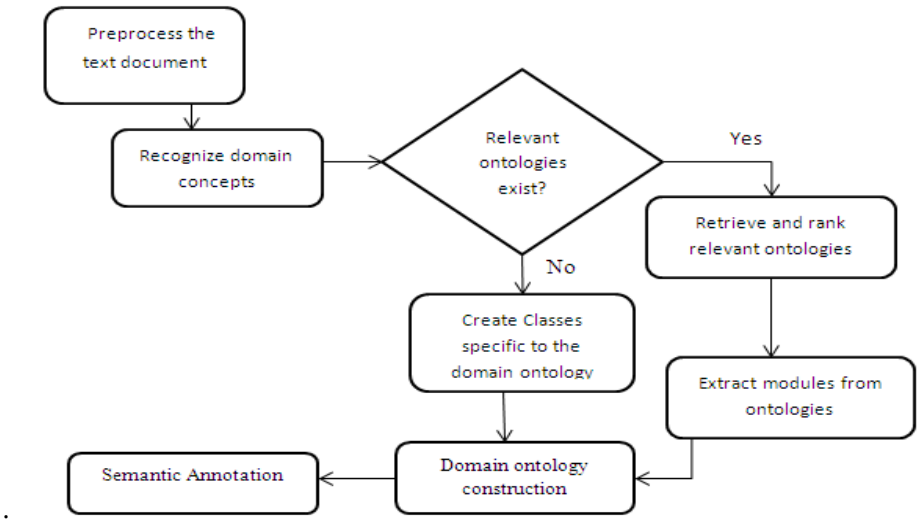

Fig. 1. Proposed system architecture.

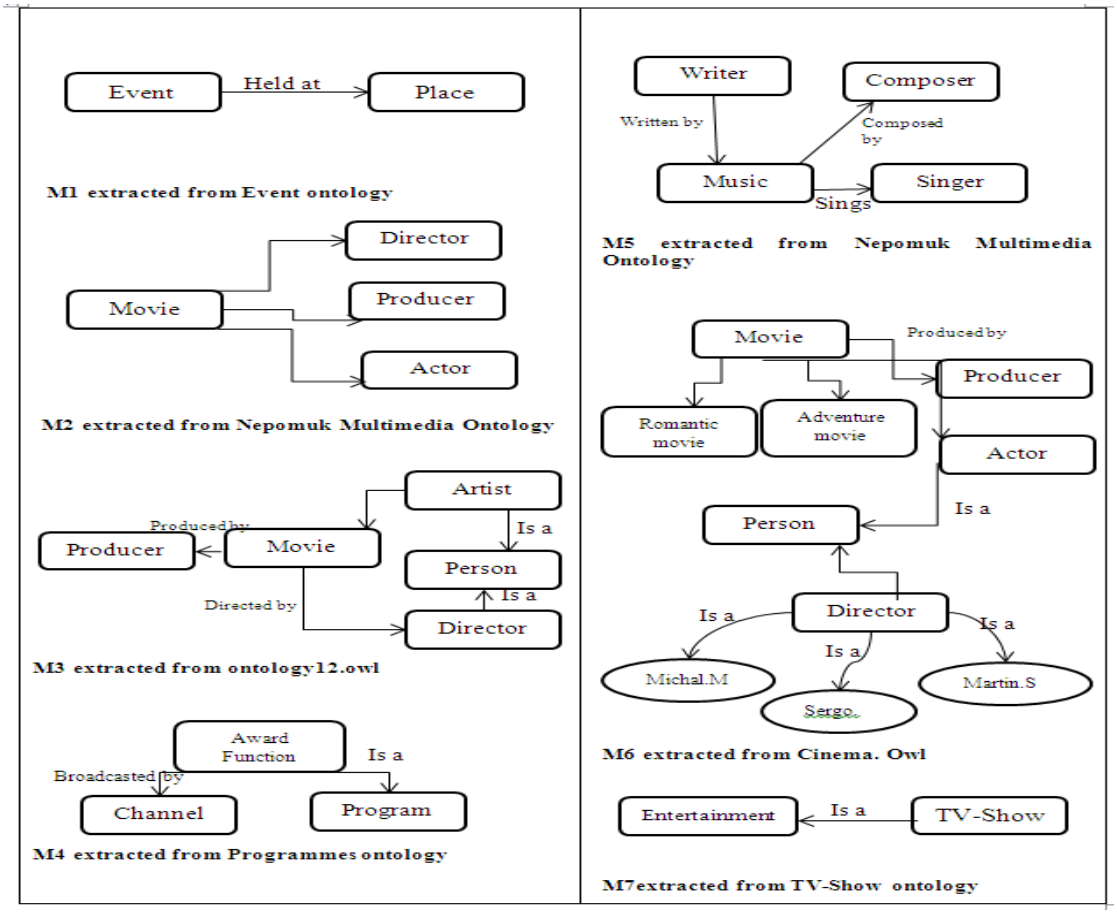

Fig. 2. Modules extracted from relevant ontologies.

B. Step 2: Searching and Ranking Relevant Ontologies

In this step the selected domain concepts are used as query to retrieve the reusable relevant ontologies from the web. We have used semantic web search engine SWOOGLE[1] as well as Google for retrieving the ontologies. The result set of ontologies obtained after searching the web is ranked according to the relevancy of ontology to the domain concepts. The ranking measure is defined as in equation 2, where O is the ontology whose rank is being calculated and k_i is the set of keywords

C. Step 3: Extracting Relevant Modules from the Ontologies

A module M from an ontology O is defined as "Minimal Portion of an ontology O that if integrated with ontology O1 yields the same effect as if the whole ontology O was integrated with O1". We have proposed an algorithm for module extraction that matches the ontology classes with the set of domain concepts WordNet [10] is used to get the synonyms of the ontology classes so that if a class label is similar in meaning to any of the domain concept is also retrieved.

Algorithm: Extracting Modules from an input ontology against a set of Domain concepts.

Input: Ontology O in OWL format, domain Concepts= K

Output: Module M

Procedure: Module Extraction (O, K)

Check_List: Container for all nodes in the ontology O.

Module_List: Container for nodes included in the extracted module.

Load the ontology

Check_List = Get all classes their relations, individuals and restrictions.

Recent= first element of Check_List.

While (Check_List! = Empty)

x = recent. Class

AccessWordNet for synonyms

S= getSyno(x)

If ((x = = any of the K OR any member of S = = any of the K) AND (x is not the member of Module_List)

Module_List. Class= Recent. Class

Module_List. Relations= Recent. Relations

Module_List. Individuals = Recent. Individuals

Module_List. Restrictions= Recent. Restrictions

End if

Recent = Next element of Check_List

End While

D. Step 4: Constructing Domain Ontology

In this step the extracted modules are assembled with the classes which are defined specifically for the desired domain ontology. Extracted modules and specifically created concepts are integrated in incremental way, thus removing any inconsistencies in ontologies parallel with the ontology construction process.

E. Step 5: Semantic Annotation

At this step the input web page is semantically annotated with the domain ontology. The domain ontology is looked up to find matching between the ontology concepts and extracted domain concepts and relationships. If an ontology concept is not directly matched to the domain concepts then subsumption relationships are used to map them.

IV. Results and Discussion

TABLE I: Extracted Relevant Ontologies

S. No	Ontology
1	Nepomuk Multimedia Ontology (NMM)
2	http://labotalc.loria.fr/~kasimir/downloads/owl/cinema.owl
3	http://tw.rpi.edu/wiki/Special:ExportRDF/Category:TV_Show.owl
4	http://raimond.me.uk/c4dm/music.owl
5	http://lsdis.cs.uga.edu/proj/traks/ontologies/ontology_12.owl
6	http://motools.sourceforge.net/event/event.html(event Ontology)
7	http://www.aktors.org/ontology/portal
8	http://www.bbc.co.uk/ontologies/programmes/2009-04-17.shtml (programmes ontology)

For the narration of proposed technique using case study, we have constructed an ontology for a document that describes the event of an award show that is stored in a local directory. Domain concepts are extracted by applying the formulas given in proposed technique for domain concept extraction. These domain concepts are then used as query to the search engine for retrieving the relevant ontologies. Set of domain concepts are stored as a set K.

K: { Entertainment, TV-show , Award Show , Award distribution, Performances , Actor, Director, Producer, Movie, Music, Nominees, Award winners, Audience , Guests }.

Retrieved ontologies are then ranked according to their degree of relevancy to the set K by using the proposed Rank measure. Table I shows the list of top ranked ontologies. By applying the proposed algorithm for module extraction, modules extracted from ontologies are shown in Fig. 2. By combining the ontology modules shown in Fig. 2 with the classes defined specifically for the "Award function ontology" the construction of domain ontology is completed. Graphical representation of the domain ontology is given in Fig. 3. Note that in the constructed domain ontology of Award Show some concepts (e.g. Nominees, Award distribution etc.) are created from scratch because these were not matched to any of the existing ontology on the web.Now by tagging the words in text document with ontology concepts make the text document machine understandable. For example if some person searches for the word Filmfare or Oscar then it is known to the search engine that it is an award function and an award function is broadcasted by some channel . By this context knowledge, achieved by ontology based annotation, the search engine will retrieve the pages which do not explicitly include the word filmfare or Oscar but they are about the award achievements of a movie or director.

[1] http://swoogle.umbc.edu/

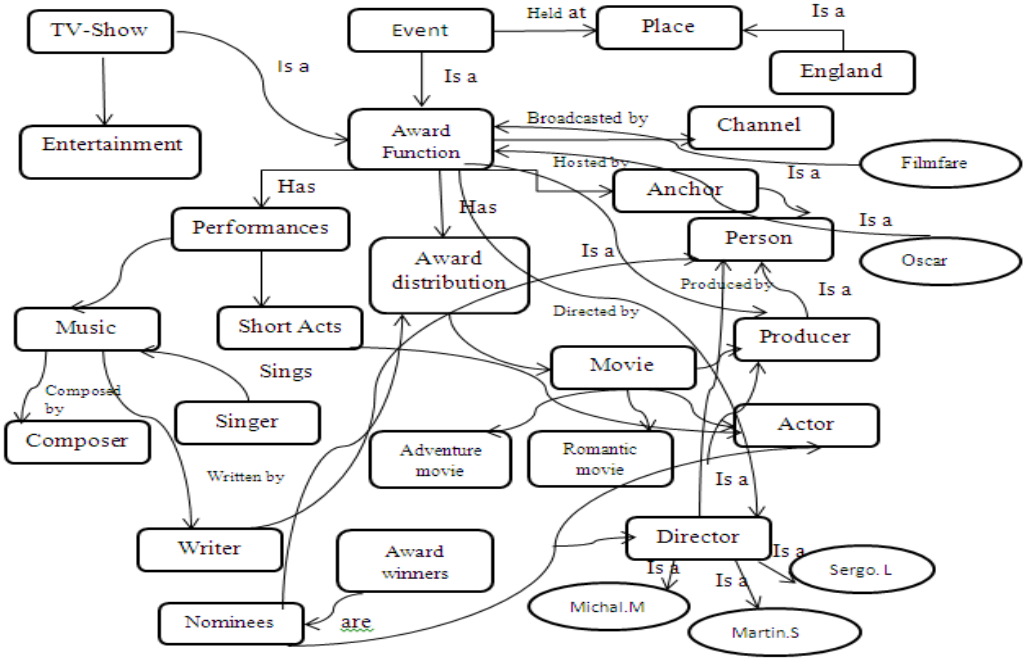

Fig. 3. Domain ontology for an award function.

V. Conclusion

There are many approaches that separately deal the issue of constructing domain ontologies and semantic annotation. But there is no complete methodology for constructing domain ontologies by partial reuse and using them to semantically annotate the text documents. This paper proposes a complete methodology for domain ontology construction for semantic annotation of text documents. Domain ontology is constructed by partially importing the existing relevant ontologies from web thus avoiding the difficulty of having to construct the ontology from scratch.

Acknowledgement

This work was supported by Higher Education Commission (HEC), Government of Pakistan under the HEC project titled: International Research Support Initiative Program.

References

[1] D. Sleeman, S. Potter, D. Robertson, and M. Schorlemmer, "Ontology extraction for distributed environment," *Knowledge Transformation for the SemanticWeb*, IOS Press, pp. 80–91, 2003.

[2] C. Brewster, F. Ciravegna, and Y. Wilks, "Background andforeground knowledge in dynamic ontology construction:viewing text as knowledge maintenance", *Semantic WebWorkshop, SIGIR'03*, 2003.

[3] H. Stuckenschmidt and M. Klein, "Structure-based partitioningof large concept hierarchies," *International Semantic Web Confernce*, pp. 289-303, 2004.

[4] B. C. Grau, I. Horrocks, Y. Kazakov, and U. Sattler, *Extracting modules from ontologies:a logic-based approach*, Springer-Verlag, pp. 159-186, 2009.

[5] F. Mostowfi, F. Fotouhi, and A. Aristar, *Ontogloss: An Ontology*-Based *Annotation Tool*, *EMELD*, 2005.

[6] J. Kahan, M. Koivunen, E. P. Hommeaux, and R. Swick, "Annotea: an open rdf infrastructure for shared webannotation," in *Proc. WWW10*, Hong Kong, pp. 623-632, May 2001.

[7] S. Staab, A. Maedche, and A. P. Handschuh, "An annotationframework for the semantic web," *in Proc. Int. Workshop onMultiMedia Annotaion*, Tokyo, 2001.

[8] F. REN, "A demo for constructingdomain ontology fromacademic papers," *in Proc. COLING*, pp. 369–376, 2012.

[9] L. Q. Troyka and J. W. Thweat, *Structured reading*, New Jersey: Prentice-Hall, 1992.

[10] G. A. Miller, "Wordnet: a lexical database forenglish," *Communications of the ACM*, vol. 38, no. 11, pp. 3941, 1995.

Tayybah Kiren was born in Pakistan, who is working as assistant professor in Computer Science Department Rachna College of Engineering and Technology Gujranwala. She is currently PhD scholar at UET Lahore under the supervision of Prof. Dr. Muhammad Shoaib.

Muhammad Shoaib was born in Pakistan, who is a professor at Computer Science and Engineering Department at the University of Engineering and Technology (UET) Lahore, Pakistan. He has completed his Ph.D. from the University of Engineering and Technology, Lahore, Pakistan in 2006. His Post Doc. is from Florida Atlantic University, USA, in 2009. His current research interests include Information Retrieval (IR) Systems, Information Systems, Software Engineering and Semantic Web.

Sang-Jo Lee was born in South Korea, who is a professor at Department of Computer Engineering Kyungpook National University; He has completed his Ph.D. from Seoul National University (SNU) South Korea. His research interests are Natural Language Processing, Information Extraction, Information Retrieval, Machine Learning, Ontology and Semantic Web.