Journal of Mathematics and System Science 3 (2013) 319-326

# Enhancing Domain Knowledge with Semantic Models of Web Documents

Anna Rozeva

*Department of Applied Mathematics and Informatics, Technical University-Sofia, 1000 Sofia, Bulgaria*

Received: May 28, 2013 / Accepted: June 26, 2013/ Published: July 25, 2013.

**Abstract:** The paper considers the problem of semantic processing of web documents by designing an approach, which combines extracted semantic document model and domain- related knowledge base. The knowledge base is populated with learnt classification rules categorizing documents into topics. Classification provides for the reduction of the dimensionality of the document feature space. The semantic model of retrieved web documents is semantically labeled by querying domain ontology and processed with content-based classification method. The model obtained is mapped to the existing knowledge base by implementing inference algorithm. It enables models of the same semantic type to be recognized and integrated into the knowledge base. The approach provides for the domain knowledge integration and assists the extraction and modeling web documents semantics. Implementation results of the proposed approach are presented.

**Keywords:** Semantic model, knowledge base, document classification, domain ontology, knowledge integration.

## 1. Introduction

The prevailing content of files and pages online and offline available in the web is textual. The amount of these resources is continuously growing and this makes the need for retrieving information by content more challenging. The extraction of semantic models of text documents and their processing with machine learning algorithms can be organized in a knowledge base. It is the means that will make the information contained in different web text resources available and provide for its efficient browsing, searching, retrieval and categorization.

Knowledge in different domains is available in the semantic web in the form of ontology. Formal knowledge structured in ontology represents a working model of entities and interactions in general or in some particular domain [1]. It includes terms with specification of their meaning, the way they are related and the constraints on their possible interpretation. This conceptualization, when formally encoded, represents the vocabulary and the semantic structure for information exchange concerning particular domain.

The analysis of unstructured or semi-structured text is performed mostly by applying machine learning techniques. For this purpose text is turned into data that can serve as input to machine learning algorithms. This is done by means of text mining [2] which extracts meaning from text in the form of terms or concepts and their relationships to documents as calculated measures of their occurrence.

The creation of semantic space of the features being input variables to machine learning provides for obtaining semantic models of the data in the analyzed domain [3]. The method of canonical correlation analysis uses complex labels for guiding the feature selection towards the underlying semantics. By means of identifying linear relationships between multidimensional variables and using two views of the same semantic object a representation of the semantics is extracted. Generation of semantic feature

**Corresponding author:** Anna Rozeva, Ph.D., associate professor, research fields: context-based text analysis, knowledge bases, semantic modeling. E-mail: arozeva@hotmail.com.

representation from a probabilistic data model suggests a way of incorporation prior knowledge of the domain and allows the use further on of different analysis methods. Both methods create semantic models by analyzing the spread of the data. Linear kernel on the term frequencies or Term Frequency Inverse Document Frequencies (TFIDF) approach is applied to learning semantics from text data. The kernel means initial projection of data in higher dimensional space and performing canonical correlation analysis in this feature space. The use of semantic models with machine learning algorithms provides for obtaining the taxonomy of text documents.

Another approach to semantic modeling concerns ontological representation of rules learnt from text by machine learning algorithm [4]. Semantic model based on rules has the advantage that irrelevant feature terms have been discarded. This semantic model is implemented in concept analysis for combining concepts into ontology classes and exploring relations between them.

Semantic model of data extracted from web pages is obtained by content-based classification that learns the data structure [5] and definitions [6]. The text is represented as a sequence of patterns containing strings with different character types. Patterns associated with a semantic type are learnt from example values of the type. They are used for the recognition of new instances of the semantic type by evaluating how well they describe the new data by comparison with known sources of information. By performing inductive logic search in the space of possible definitions the semantic model best fitting for each new source is learnt.

For the purpose of making automated use of text web documents the aim of the current research is to define a framework for their semantic modeling and for aligning the model to existing domain knowledge. The paper is organized as follows: Section 2 contains description of the designed framework for semantic modeling. Section 3 considers method for aligning the model to domain knowledge represented by ontology. Section 4 presents results of the framework implementation on extracted web documents. Section 5 concludes the work with discussion of the results obtained and directions for future work.

## 2. Semantic Model of Web Text Documents

Semantic modeling for providing semantic access to web resources is implemented by creating an RDF model that references HTML fragments by means of vocabulary of terms and their mapping to RDF resources [11]. Semi-automatic semantic document modeling by implementing domain ontology on HTML document resulting in XML representation of the concepts and relationships in the documents is presented in Ref. [12]. Another approach for semantic modeling is ontology based and is implemented as network of concepts with explicit relationships [10]. Adding more individuals for supporting individual reasoning extends the model. Semantic model representing relational database obtained from ontology input [16] is designed with the aim of enhancing the efficiency of answering semantic queries about ontology instances. Architecture for semantic modeling implementing lexical database and text engineering architecture for ontology population with XML files is shown in Ref. [8]. Context modeling of text based on the vector space model as the common model in natural language processing and information retrieval is implemented in Ref. [15]. Approach for selecting appropriate context units carrying semantic information and defining the appropriate strength weights for discriminating the meaning of text in the form of frequent patterns mined from it is proposed. The semantic analysis is performed by means of semantic similarity measure computed by cosine function of context vectors. Distributional model of semantics based on matrix factorization [17] yields semantic representation of words that is claimed to be sparse, effective and highly interpretable.

The approach proposed for designing semantic model of text proposed in the current paper represents

the taxonomy generated from the retrieved documents. It is learnt by classification of the semantically labeled documents' feature space. This is inspired by the fact that classification is the most natural way for enabling the interpretation of unstructured and semi-structured text referring to particular domain. The goal is from the set of retrieved documents referring to a given domain (1) to design a structure representing the document descriptive terms and (2) to learn rules of their grouping and relations. The rules, represented as taxonomy, will enable reasoning and inference for obtaining the most relevant semantic model of the web documents. The proposed framework for extraction of web text documents semantic model and its alignment to domain knowledge represented by text taxonomy is shown in Fig. 1.

Text retrieved from web documents is processed for feature selection. Features represent terms (words) and the task is to extract the ones best distinguishing the document content. Eq. (1) presents the TFIDF [3] measure for term selection.

$$TFIDF(d_i w_j) = Nw_j d_i \times \log \frac{Nd}{Nw_j} \quad (1)$$

The notation used in Eq. (1) is as follows: di is a text document, wj is term, Nwjdi is the number of occurrences of the term in the document, Nd is the number of documents and Nwj is the number of documents that contain the term wj. Eq. (2) represents the obtained document feature space.

$$FS = \{d_i[(w_1, TFIDF_1), (w_2, TFIDF_2), \dots, (w_n, TFIDF_n)]\} \quad (2)$$

The feature space described by Eq. (2) is considered to be insufficient for creating domain specific semantic model of web text. Semantic enrichment by assigning semantic annotations to the features with links to their semantic descriptions is considered in Ref. [14]. The metadata provide class and instance information. The prerequisites for representation of semantic annotations defined there are: ontology defining the entity classes, entity identifiers for linking to the semantic descriptions and a knowledge base with entity descriptions and relationships. The method proposed in the framework in Figure 1 performs the enrichment of

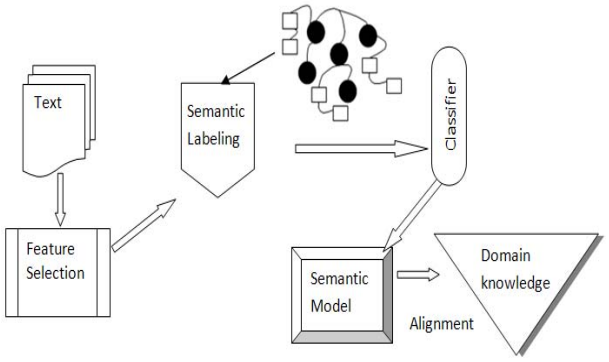


**Fig. 1 Framework for semantic modeling of web text and alignment to domain knowledge.**

the feature space with domain related semantics by implementing domain ontology. It provides for adding semantic labels to the features in the feature space. The labels represent either ontology classes or individuals. The semantic labeling ensures the conceptualization and identification of the feature space. Object properties defined in ontology are considered and features are mapped against them. The mapping is implemented by description logic queries. Ontology querying is discussed in Ref. [9]. Graphical tool for building queries has been designed and the queries are submitted to a reasoner. Application for querying OWL ontologies is presented in Ref. [7]. It makes use of the built-in query engine of a reasoner. The method proposed in the current paper implements query agent for queries' instantiation and execution with expected result that asserts features as ontology individuals or classes. Features that have not been mapped to ontology entities will be analyzed at the stage of aligning the semantic model to domain knowledge base. The algorithm for performing the semantic labeling is shown in Fig. 2.

The method performs focused semantic labeling considering the goal for extracting domain related text documents.

The features in the semantic feature space are processed further on by classifier for learning rules that expose dependencies and relationships between them. Assignment of classification topics is performed by the method presented in Ref.[13]. The rule space generated consists of rule paths RP, described by Eq. (3).

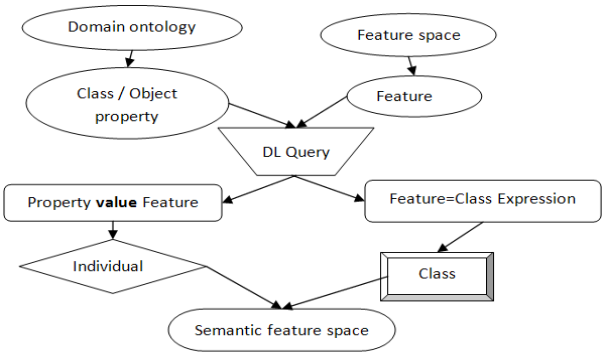


**Fig. 2 Generation of semantic feature space**

$$RP = \{[(w_1, st_1, s_1, p_1), (w_2, st_2, s_2, p_2), \ldots, (w_n, st_n, s_n, p_n)]\} \quad (3)$$

A rule path consists of elements that contain a feature wi , state sti (Boolean), semantic label si and a probability measure pi. If the label represents individual it includes the class as well. The taxonomy of semantically labeled rules represents the resultant semantic model of the input web documents.

## 3. Semantic Model Alignment to Domain Knowledge

Domain knowledge shown in the framework from Figure 1 represents a rule knowledge base. The rules R have the form:

∪Condition(feature, state, label, probability) →Conclusion.

It has been populated from trusted sources and conforms to existing domain ontology. The semantic model of retrieved web text is to be aligned to the knowledge base. The alignment concerns adding rules, which contain new knowledge and discarding the ones, which are subsumed by, similar to or mutual exclusive with existing rules. These relations have been defined and examined in the context of learning ontologies from rules in Ref. [4]. Rules from the semantic model are to be compared to ones in the knowledge base that imply one and the same conclusion. Fig. 3 presents a method designed for the semantic model alignment to the knowledge base.

The description of cases providing for adding or discarding semantic model rules uses the following notation: SR – rule of the semantic model; F – feature

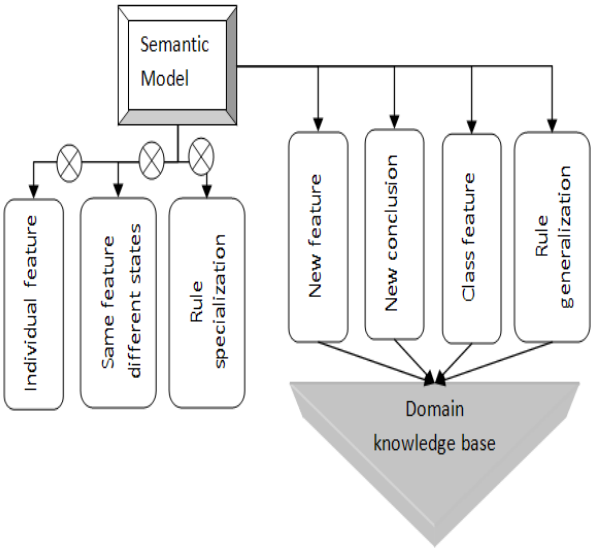


**Fig. 3 Semantic model alignment to a knowledge base.**

that is part of rule path condition CD; L –semantic label representing ontology class; LI – ontology individual label; S – feature state; N – number of rules in the base having the same conclusion C; M – number of conditions in a rule path.

New feature concerns rules whose path involves condition on semantically labeled feature not matching any of the ones in existing rules. Adding them to the base provides for the enrichment of its feature space. Eq.4 gives new feature case description.

$$\exists F_{SR} \notin \{F_{i,j}, i = 1, M; j = 1, N\} \wedge L \neq \emptyset \quad (4)$$

Rules from the model resulting in new conclusions currently not present in the base will be included as well. Eq. (5) presents the new conclusion case description.

$$SR(C) \notin \{C_i, i = 1, N\} \quad (5)$$

Class rule (Eq. (6)) refers to the case when rule from the model involves condition on feature semantically labeled with ontology class while the existing rule condition is labeled as individual from that class if the model rule probability is higher.

$$M_{SR} > M_{Ri}, i = 1, N \quad (6)$$

Rule generalization represents the case when the semantic model rule path involves more conditions compared to existing rules in the knowledge base. Adding such rules to the base provides for the enhancement of its semantics with feature relationships.

Eq. (7) gives the case definition.

$$M_{SR} > M_{Ri}, i = 1, N \quad (7)$$

The method for aligning generated semantic model to domain knowledge base considers a rule as not suitable for enhancing the base if it involves condition with feature state opposite to the state of the same feature in existing rule. Such rules are considered mutual exclusive with case definition presented by Eq. (8).

$$F_{SR} = F_{Ri} \wedge S(F_{SR}) \neg S(F_{Ri}) \quad (8)$$

If condition in a rule path has feature that is semantically labeled as ontology individual and there is a rule in the base involving feature referring to the individual's ontology class then these rules are considered similar. Rule similarity definition is analogous to the one given by Eq. (6) with inclusion relation '⊂' between feature labels.

Rules in the model involving same terms but fewer conditions than existing ones are considered subsumed and are not added to the knowledge base. Definition of rule subsumption is analogous to that in Eq. (7) with '<' relation between the number of conditions in the corresponding rule paths.

## 4. Framework Implementation and Results

### *4.1 Framework Set Up*

The framework proposed in the current paper has been implemented on web text documents of conference proceedings on e-Governance extracted from (http://fman.tu-sofia.bg). The document feature space has been initially obtained. The feature space labeling for generating the semantic model has been performed by using e-Governance ontology presented in Ref. [18] and shown in Fig. 4.

### *4.2 Semantic Labeling*

Semantic labeling is performed in Protégé [19] by query agent executing logic queries involving terms from the feature space against the domain ontology. Initially features are used as query input for retrieving ancestor classes or individuals. Sample query with the feature "E-Governance-Structure" as class expression is shown in Fig. 5.

If the query result involves ancestor classes or individuals the feature is labeled as ontology class:

(E-Governance-Structure, class)

If the query result is empty, i.e. the feature doesn't represent an ontology class, query for retrieving individuals is designed. It involves object property and the feature examined as predicate value. The output is set to superclasses and individuals. The feature is checked with all object properties until the query produces a non-empty output. Sample query with object property "hasTermValue" and the feature "Administrative" is shown in Fig. 6.

If the superclass result set is non-empty and the query output involves individuals, the feature is semantically labeled as individual:

(e-government, individual)

In case that the feature matches neither class nor individual it is labeled as "null".

Learning the semantically labeled feature space with decision tree classifier generates the semantic model. It involves rules obtained from the tree nodes. Sample rule path, produced in Microsoft SQL Server [20] by the classifier is:

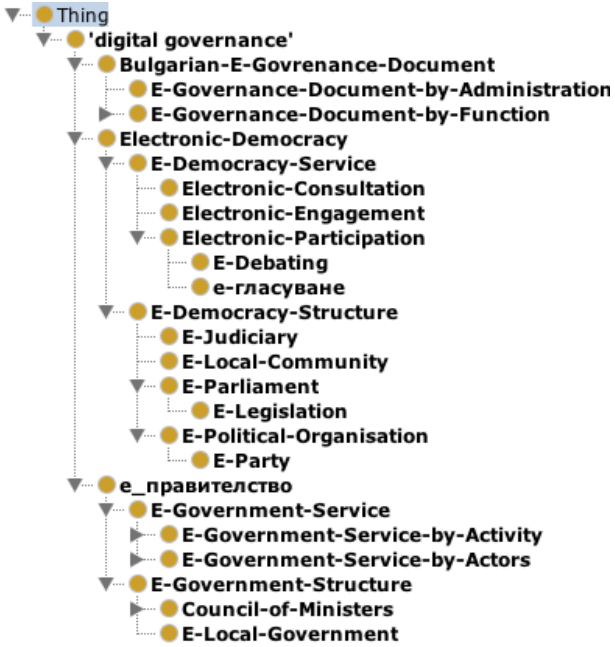


**Fig. 4 An excerpt of e-Governance ontology.**

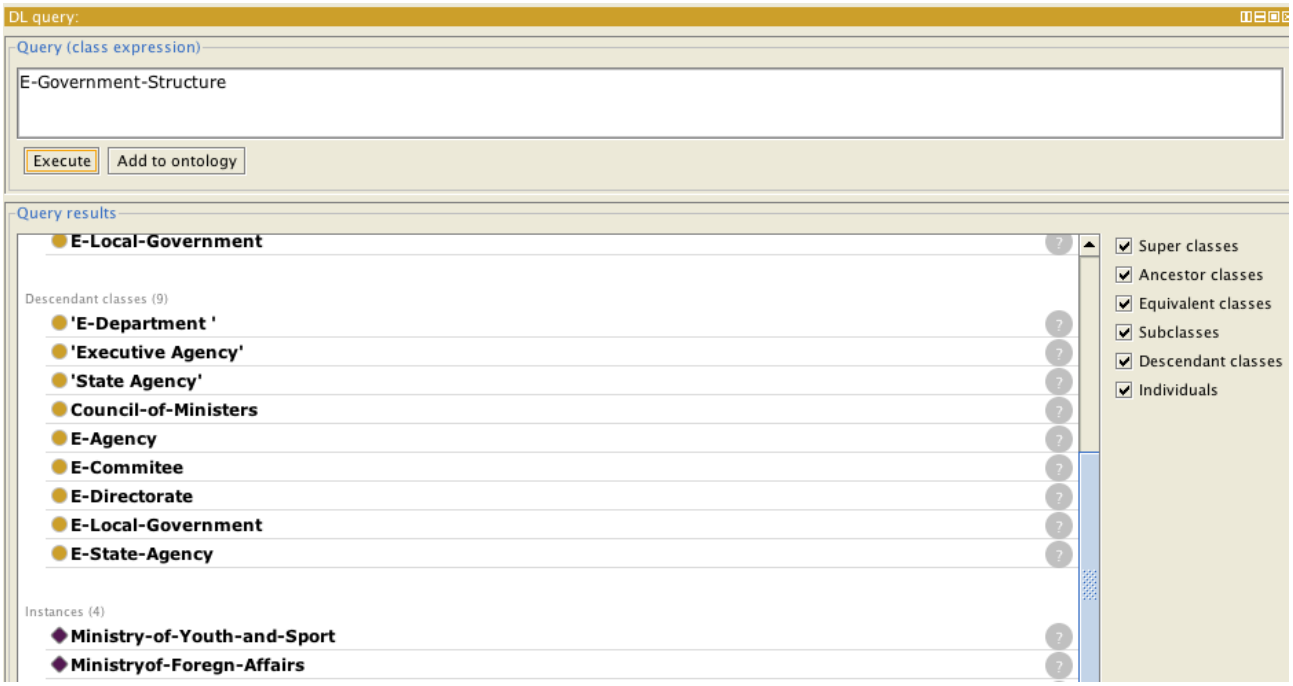


**Fig. 5 Logic query with class expression.**

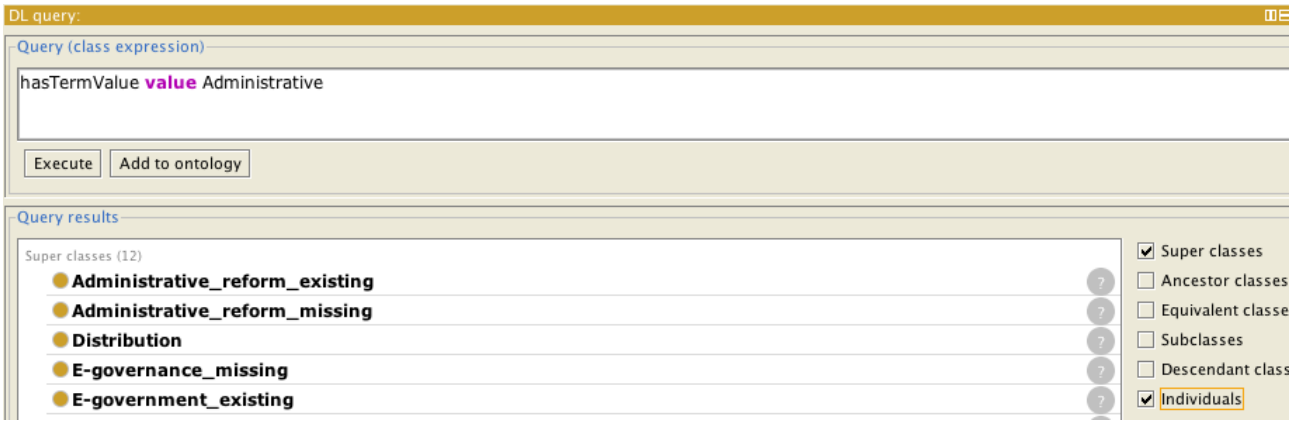


**Fig. 6 Logic query with object property and the feature as predicate value.**

E- Gov 65 Vectors(Business Process, null) = Missing

and E- Gov 65 Vectors(environment, null) = Missing

and E- Gov 65 Vectors(e-governance, class) = Missing

and E- Gov 65 Vectors(e-Government, class) = Missing

and E- Gov 65 Vectors(Public administration, null) = Missing

and E- Gov 65 Vectors(result, null) = Missing

"Missing" is replaced by "0" and "Existing" - by "1" in the model. Rule conclusions represent the distribution of documents into classification categories, shown in Fig. 7.

### *4.3 Semantic Model Alignment to Domain Knowledge Base*

The semantic model of new documents containing

| ATTRIBUTE_NAME | ATTRIBUTE_VALUE | SUPPORT | PROBABILITY |
|---|---|---|---|
| Topic | Missing | 0 | 0 |
| Topic | A | 0 | 0,0833333333333333 |
| Topic | B | 1 | 0,166666666666667 |
| Topic | C | 7 | 0,666666666666667 |
| Topic | D | 0 | 0,0833333333333333 |

**Fig. 7 Semantic model rule conclusion.**

**Table 1 Enhancement of e-Governance domain knowledge with semantic model of web text documents.**

| Semantic model rule | Knowledge base rule | Case | Action |
|---|---|---|---|
| 1: (business process, 0, null), (communication, 1, null) | | New conclusion (new classification topic E) | Add rule |
| 2: (e-governance, 0, class), (e-government, 1, class), (public administration, 0, public governance individual) | (e-governance, 0, class), (e-government, 1, class), (public administration, 0,) | Subsumption | Discard rule |
| 3: (Bulgarian e-government, 1, e-government individual), (public administration, 0, public governance individual) | (e-government, 1, class), (public administration, 0, public governance individual) | Similarity | Discard rule |
| 4: (e-government, 0, class), (public administration, 0, public governance individual) | (e-government, 1, class), (public administration, 0, public governance individual) | Mutual exclusivity | Discard rule |
| 5: (e-governance, 0, class), (e-government, 1, class), (public governance, 0, class) | (e-governance, 0, class), (e-government, 1, class), (public governance, 0, individual) | Class rule | Add rule |
| 6: (e-governance, 0, class), (e-government, 1, class), (public administration, 0, public governance individual), (E-Directorate, 1, class) | (e-governance, 0, class), (e-government, 1, class), (public administration, 0, public governance individual), | New feature | Add rule |
| 7: (business process, 1, public governance individual), (environment, 0, null), (e-governance, 0, class), (e-government, 1, class), (public administration, 0, public governance individual) | (e-governance, 0, class), (e-government, 1, class), (public administration, 0, public governance individual) | Rule generalization | Add rule |

the described rule paths has been aligned to e-Governance domain knowledge base. The knowledge base has been populated with rule paths, extracted from the same trusted source. The knowledge base rules have been labeled with entities of domain ontology. Sample rule path from the base is:

(e-governance, 0, class), (e-government, 1, class), (public administration, 0, public governance individual), (result, 0, null), (aim, 1, null).

The enhancement of the existing knowledge base with the semantic model according to the defined method is performed by matching model and knowledge base rules initially according to their conclusions and further on the rules having the same conclusions. Examples of cases described in the method and the actions performed for enhancing the existing domain knowledge are shown in Table 1.

## 5. Conclusions

The semantic modeling of text retrieved from web resources is designed in a framework which involves text mining techniques, ontology querying for extracting semantic labels of the feature space and machine learning for reducing the dimensionality of the feature space and discovering relations among the features. Ontology querying algorithm is designed which outputs classes or individuals as feature labels. Semantic model of this type provides for maintenance and integration with existing domain knowledge bases. The proposed method defines conditions for enhancing a knowledge base by analyzing rules providing new or similar conclusions. The analysis asserts mutual exclusive conditions in rule paths from the Boolean attribute for feature presence or absence in the text, rule subsumption from the path length and similarity from the semantic labels. Implementation of semantic model of e-Governance text documents and ontology and its integration with existing domain knowledge base is presented.

Future work is intended in designing method for consideration of features with missing semantic labels as well as the role of the probability attribute in the knowledge integration process.

## References

[1] R. Stevens, C. A. Goble, S. Bechhofer, Ontology-based Knowledge Representation for Bioinformatics, Brief Bioinform 1 (4) (2000) 398-414.

[2] J. Clark, Text Mining and Scholarly Publishing, Publishing Research Consortium [Online], February 2013, http://www.publishingresearch.net/documents/PRCText MiningandScholarlyPublishinFeb2013.pdf, (accessed May 10, 2013).

[3] D. R. Hardoon, Semantic Models for Machine Learning, Ph.D. Thesis, University of Southampton, 2006.

[4] H. Suryanto, P. Compton, Discovery of Ontologies from Knowledge Bases, in: International Conference on Knowledge Capture, New York, NY, USA, ACM Press, 2001, pp.171-178.

[5] J. L. Ambite, B. Gazen, C. A. Knoblock, K. Lerman, T. Russ, Discovering and Learning Semantic Models of Online Sources for Information Integration, in: IJCAI Workshop on Information Integration on the Web, Pasadena, CA, 2009.

[6] M.J. Carman, C.A. Knoblock, Learning semantic definitions of online information sources, Journal of Artificial Intelligence Research 30 (2007) 1-50.

[7] G. Tzoganis, D. Koutsomitropoulos, T.S. Papatheodorou, Querying ontologies: Retrieving knowledge from semantic web documents, in: Proc. of the 3d Panhellenic Student Conference on Informatics, Related Technologies and Applications, 2009.

[8] H. Alani, S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis, N. R. Shadbolt, Automatic Ontology-Based Knowledge Extraction from Web Documents, IEEE Intelligent Systems, (2003) 14-21.

[9] A. Fadhil, V. Haarslev, OntoVQL: A Graphical Query Language for OWL Ontolgies, in: D. Calvanese, E. Franconi, V. Haarslev, D. Lembo, B. Motik, S. Tessaris, A.-Y. Turhan (Eds.), Proceedings of the 20th International Workshop on Description Logics DL'07, 2007, pp.267-274.

[10] W. Li, C. Yang, R. Raskin, A Semantic Enhanced Model for Searching in Spatial Web Portals, in: D. McGuinness, P. Fox, B. Brodaric (Eds.), Semantic Scientific Knowledge Integration Papers from the 2008 AAAI Spring Symposium Technical Report SS-08-05, 2008, pp. 47-50.

[11] J.I. Fernandez-Villamor, J. Blasco-Garcia, C.A. Iglesias, M. Garijo, A Semantic Scraping Model for Web Resources – Applying Linked Data to Web Page Screen Scraping, in J. Filipe, A.L.N. Fred (Eds.), ICAART 2011 - Proceedings of the 3rd International Conference on Agents and Artificial Intelligence, Volume 2 – Agents, 2011, p.451-456.

[12] S. A. Noah, L. Zakaria, A. C. Alhadi, Extracting and Modeling the Semantic Information Content of Web Documents to Support Semantic Document Retrieval, in: Proceedings of the Sixth Asia-Pacific Conference on Conceptual Modeling, Vol.96, 2009 pp.79-86.

[13] A. Rozeva, Classification of text documents supervised by domain ontologies, ATI-Applied Technologies & Innovations 8 (3) (2012) 1-12.

[14] A. Kiryakov, B. Popov, I. Terziev, D. Manov, D, Ognyanoff, A. Kirilov, M. Goranov, Semantic Annotation, Indexing and Retrieval, Web Semantics: Science, Services and Agents on the World Wide Web 2 (1) (2004) 49-79.

[15] Q. Mei, D. Xin, H. Cheng, J. Han, C.X. Zhai, Generating Semantic Annotations for Frequent Patterns with Context Analysis, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06), 2006 pp. 337-346.

[16] P. Lependu, D. Dou, G.A. Frishkoff, J. Rong, Ontology Database: A New Method for Semantic Modeling and an Application to Brainwave Data, B. in: SSDBM'08 Proceedings of the 20th International Conference on Scientific and Statistical Database Management, 2008, pp. 313-330.

[17] B. Murphy, P.P. Talukdar, T. Mitchell, Learning Effective and Interpretable Semantic Models Using Non-Negative Sparse Embedding, in: Proceedings of COLING 2012: Technical Papers, 2012 pp.1933-1950.

[18] B. Deliyska, R. Ilieva, Ontology-Based Model of E-Governance, Annual of "Informatics" Section, Union of Scientists in Bulgaria, 4, (2011) pp.103-119.

[19] Protégé Home Page, http://protege.stanford.edu (accessed May 10, 2013).

[20] Microsoft SQL Server, Home Page, http://www.msdn.microsoft.com/en-us/sqlserver/default (accessed May 10, 2013).