Classification of text documents supervised by domain ontologies

Anna Rozeva
Faculty of Business Management, University of Forestry, Bulgaria

The research objective is to establish an approach for supporting the classification of text documents referring to a specified domain. The focus is on the preliminary topic assignment to the documents used for training the model. The method implements domain ontology as background knowledge. The idea consists in extracting the preliminary topics for training the classifier by means of unsupervised machine learning on a text corpus and further alignment of the document vectors to concepts of the ontology. The results obtained by classification of new documents supervised by e-governance ontology with several machine learning algorithms showed sufficient match of their content to the ontology concepts. A conclusion is drawn that the approach can support the automatic extraction of documents relevant to any domain described by ontology.

Keywords: Text classification, Topic assignment, Supervised learning, Ontology, E-governance

Introduction

The problem of automatic classification of texts has been intensively researched in the recent decade. The demand for supporting tools and methods is increasing with the continuously growing amount of texts available in e-libraries, web pages, blogs, forums, etc. On the other hand having big text volumes available the need for explicitly classifying texts as pertaining to specifically selected subject domain is also an extremely important topic in information processing technologies for knowledge discovery. This task can be facilitated by providing means for supervising the classification of the text corpus examined for matching the domain of discourse.

So far as ontologies have emerged as the means for capturing knowledge from an existing domain, application, task, etc. it is most reasonable to involve them in the classification process for supporting semantic-based applications by providing the required domain-oriented conceptualization. With the growth of the Semantic Web and the presence of a lot of tools for ontology creation, management and sharing, ontologies concerning various knowledge domains are already available. Their involvement into machine learning-based knowledge discovery has increased and promising results have been obtained.

Current work aims to contribute to the automatic explicit text classification by implementing ontologies. The initial assignment of topics to the training documents of the classification model is proposed to result from unsupervised machine learning which are further on aligned to ontology concepts. This method results in tuning the classifier to the background knowledge for the selected domain. The enhanced classifier is to provide for more trustworthy explicit classifications of incoming texts. The approach has been implemented on texts concerning e-governance coupled with an existing e-governance domain ontology.
Related work

The review of related work will concern achievements in the text classification with ontology support on the one hand and state of the art of ontologies pertaining to the domain of e-governance.

Approaches for text classification involving ontologies

Architecture of text classification system with ontological, structural and categorization layers is presented in (He, Qui, Zhao, and Wang, 2004). The ontology layer involves a couple of ontologies of the domain, i.e. structure, domain and vocabulary. The structure ontology provides for the extraction of structured items from the text which by means of the vocabulary provides them with the related domain knowledge. The ontology layer assists the structural one for producing well structured output documents that form a vector space with defined semantics. In order to avoid commonly encountered ambiguity of semantics, ontological support is provided as reasoning rules. The ontology refers to staff in an organization’s department and is taken from an open web site. Two machine learning algorithms are implemented in the architecture for building the classifier and for performing categorization of unknown text documents, i.e. Naïve Bayes and Ripper.

The text categorization task generally involves building a classifier with training documents and its further implementation on new incoming documents. Approach that doesn’t require e set of initially classified documents for training the classifier appears in (Janik and Kochut, 2008a). Instead ontology that contains the entities from the examined domain, the categories they belong to and the relationships among them are used as a classifier for the concepts and categories of the domain. Documents are processed for obtaining a thematic graph of entities. It is analyzed for determining its categorization according to the taxonomy of ontology categories. In order to apply the ontology as classifier the text document is to be converted into a graph. This process involves performing entity matching and identification of relationships. The classification consists in measuring the semantic similarity between the graph and the ontological categories. The approach relies on ontology with a rich instance base of interconnected entities. It is implemented on news articles with ontology derived from the English version of Wikipedia. This ontology has been chosen because of the literals and instance entities contained form a highly connected graph of statements. Further on the development of the algorithms for the semantic graph construction, the thematic graph selection and analysis and the dominant thematic graph categorization are shown in (Janik and Kochut, 2008b). The classification method selected was Naïve Bayes. The categorization has been performed both in the case of documents trained from the ontology and the ones trained from the original corpora.

Another approach for implementing ontologies for text classification and clustering tasks is considered in (Bloehdorn, Cimiano, and Hotho, 2006). The ontological structures are obtained from the text corpus itself. This is achieved as a result of unsupervised learning process. The manually engineered and automatically constructed ontology provides for augmenting the text representation as bag-of-words. Concept hierarchies are derived from the text by means of clustering techniques. The model adopts the vector-space representation of texts but enriches term features with syntactic dependencies besides the word co-occurrence. Similarity between terms is calculated and concept hierarchy is built by using clustering techniques. The analyzed text corpus represents journal article titles and abstracts, which are indexed with descriptors from manually engineered ontology.
Automatically acquired domain ontology for supporting text categorization has been used also in (Wu, Tsai, and Hsu, 2003). Ontologies are generated by applying morphological rules and statistical methods. The quality of the generated domain ontology has been tested experimentally. It is claimed that with good domain ontology the concept structure of the sentences can be identified. By compiling the concepts from documents in the training set they can then be implemented on the testing set for classification. The implemented domain ontology has been designed to integrate linguistic, common-sense as well as domain knowledge and performs natural language understanding, i.e. the event structure of sentences is identified by matching words to ontology concepts and relationships.

Categorization of document abstracts by using the taxonomy of categories in the International Patent Classification as well as in a huge number of classified patent documents has been proposed in (Seddiqui and Aono, 2009). The approach implemented in the system generates mapping of the terms in the patent documents and the preliminary classified patent classification for retrieving probable classifiers related to the examined document or abstract. Further on similarity between the probable classifier and neighboring ones is looked for. Thus the classifier is refined by producing relevant classifications to a sufficient depth in the ontology taxonomy by implementing techniques for ontology alignment.

The approaches and methodologies presented so far deal with monolingual texts. The one described in (de Melo and Siersdorfer, 2007) implements ontologies for multilingual text classification. It concerns linguistically motivated representation of text that can use classifications learnt from preliminary classified examples in one language in order to automatically classify documents provided in a different language. In using ontologies and lexical resources it goes beyond the mapping of document terms to ontological concepts and examines mapping to regions of concepts. Algorithm for exploring related concepts that might be relevant has been applied. The experiments have been made on the Reuters corpus and the WordNet 2.1. for the English language and mapping tables to the Spanish WordNet.

Approach for multi-label classification of text is proposed in (Vogrinicic and Bosnic, 2011). It implements semi-automatic generation of ontology for detecting the most notable concepts in the domain that explored documents refer to and consequently these concepts are used as possible topics for the texts whose content is not known in advance. The multi-label classification task is transformed into one or more single-label tasks. The transformation approaches can be used with an arbitrary classifier that outputs a probability distribution over classes. Implementation has been made with support vector machines, decision tree, Naïve Bayes and k-nearest neighbor for the evaluation of the transformation approaches. By direct handling of multi-label data the consideration of correlations between different labels has been achieved. Multi-label neural network and multi-label k-nearest neighbor algorithms have been used.

The related work on text classification supported by ontologies thus presented is the background for the investigation of the current paper. Its aim is to implement the classifier to a document corpus concerning the e-governance domain. The selection of ontology that will be implemented for supervising the classification process has been achieved on the basis of a review of the resources presented further on.

**E-governance ontologies**

E-governance and e-government projects have been researched and developed intensively in the last decades for facilitating service delivery to the public. Domain ontology involving semantic content from the government service domain is
designed in (Fonou-Dombeu and Huisman, 2011). The domain represents project monitoring. The concept Government is subconcept of Project and its subconcepts are government organization and municipality.

Project related ontology with knowledge about e-government project types is presented in (Sarantis and Askounis, 2010). Out of competency questions referring to types of e-government projects and the way they relate to e-government dimensions the following concepts have been defined: beneficiary, administration level, domain sector, function and nature. The administration level subsumes the national, regional and local levels.

Ontologies concerning e-government services are designed in (Apostolou, Stojanovic, Lobo, Miro, and Papadakis, 2005) and (Vassiliakis and Lepouras, 2006). The service modeling in ontology achieved in (Apostolou et al., 2005) is by meta-ontology cluster, consisting of ontologies like:

- Legal - defines legal document structure
- Organizational - defines units, role persons, resources
- Lifecycle - instances of all decisions relevant for the service
- Domain - specific domain knowledge
- Service - elements from the domain ontology as inputs and outputs and from the lifecycle ontology for explaining reasons, motivating the decisions
- Life event - models services’ categorization.

The approach for building the e-government public services ontology in (Vassiliakis and Lepouras, 2006) is based on a template of approved concept and relationship types. The concepts included are service, service consumer, organization, service implementation, legislation, form, document and life event.

Ontology scheme of e-governance consisting of meta-ontology, domain ontology, application ontologies and task ontologies of e-governance services has been found in (Deliyska and Ilieva, 2011). The e-governance taxonomy has been defined, as well as synonyms and near synonyms of classes, instances and attributes. The domain ontology of e-governance and e-government service class taxonomy and the taxonomy of specified service are shown. This ontology has been selected for the supervised text document classification as fitting best to the document corpus content.

Approach for text categorization supervised by ontology

The categorization of documents from a text corpus is performed by a classifier, obtained from a training set of documents with preliminary assigned topics (categories). An approach and modeling framework of this process have been discussed in (Rozeva, 2011) and (Rozeva, Ivanov, and Tsankova, 2011). The topics there have been manually assigned to the documents. The initial topic assignment required in the process of developing a supervised classification model by means of applying machine learning technique to the training set is generally done intuitively. Consequently the resultant classifications of new incoming texts often do not align well to the knowledge domain they are supposed to concern. The approaches implemented for converting the text to a structure applicable for processing by machine learning algorithm perform mostly word and phrase recognition and their weighting. Therefore they lack the semantics of the knowledge domain the documents refer to. Classification with an ontological classifier will provide for the discovery among incoming documents the ones that are relevant to a selected knowledge domain. Thus the ontological supervision of
the classification process provides for the identification of pieces of knowledge in the text document pool pertaining to the domain of interest. In this way the supervision of ontology adds the semantics and context to the general knowledge discovery process. The goal of the proposed approach is to relate documents’ descriptive terms to elements of specific domain knowledge available in ontology. The idea is the ontology content to guide and direct the classification process. By coupling the classifier with ontology classifications that are most relevant to the examined domain are to be obtained. The flowchart of the processing tasks is shown in Figure 1.

The ontology is presented as concept taxonomy and vocabulary. The steps of the process are as follows:

- **Term conceptualization task** - performs the transformation of terms extracted from documents to ontological concepts or instances. This is achieved by a mapping procedure to the ontology vocabulary. The vocabulary contains the ontology concepts and their synonyms. The procedure transforms terms to concepts by direct mapping or through a synonym. Terms that fail the mapping procedure are discarded. The procedure outputs ontologically conceptualized terms. They define the new document vector space dimensionality, which is significantly smaller than the initial one. Document vectors in this space are denoted as *concept vectors*.

- **Unsupervised group identification** - trains a model that will serve for the topic assignment. Document groups are identified by applying unsupervised machine learning algorithm to the training set of document concept vectors with unknown content. Groups containing similar documents are presumed to represent the source for topic identification.

- **Topic assignment** - performs topic definition by mapping the elements of the concept vector of each group member to their subsuming concept in the ontology taxonomy. Vectors whose elements do not match ontological concepts are ignored and the vector is excluded from the training set as
irrelevant to the selected knowledge domain. The topic for the group is obtained by smoothing group elements’ topics to their subsuming concept. The topic obtained by the ontology mapping procedure is assigned to the vectors in the group. By following this procedure some of the initially identified groups may produce the same ontological topics. Consequently the final number of topics assigned to the training set will decrease. Therefore in order to provide for the better refinement of the topics, the learning algorithm which produces the greatest number of groups is to be preferred.

- **Supervised training** - the set of conceptualized document vectors with ontological topic assigned is trained by classification algorithm and the resultant classification model is referred to as *ontological classifier*.

The conceptualized vectors of new documents with missing topic represent the input for the ontological classifier.

**Application of ontological text classification**

In order to apply the proposed approach a text corpus has to be prepared as well as the relevant domain ontology.

**Figure 2. Excerpt of e-Government service ontology**

Text corpus description

The text corpus examined is taken from the e-library described in (Tsankova and Rozeva, 2011) and (Tsankova, 2010). It has been established for collection and dissemination of good practices in administrative processes management. The
subject matter of the texts included therein covers state and local administration, e-solutions in the administrative practice, business processes and e-solutions in business management. It is available from http://fman.tu-sofia.bg. The documents contain title, author, abstract and content. The document corpus that has been extracted for designing the classification model involves 65 document titles and abstracts. Selected documents have been pre-processed and inserted into a table with identifier, text stream and topic columns. The topic column is initially empty.

**Domain ontology selection**

The subject of e-library documents refer to the e-government domain. Ontology describing it has been found in (Deliyska and Ilieva, 2011). An excerpt of it is shown in Figure 2.

It has been implemented for the supervised categorization of texts which are meant to refer to the e-government subject area. The ontology has been pre-processed and the concept vocabulary has been established.

**Document term conceptualization**

After the methodology of term extraction presented in (Rozeva et al., 2011) and for the sake of discriminating common sense words in the text corpus, phrases consisting of 2 words with minimum occurrence of 2 and weighted by the TFIDF (term frequency inverted document frequency) measure have been extracted. This resulted in 128 phrases scored from 6.93 to 36.72. This threshold was accepted because of the great deviations of the number of terms with occurrence frequencies 1 and 3, which were 711 and 43 respectively. The extracted terms were mapped to the ontology vocabulary resulting in 87 conceptualized terms. The complement to the initial number represents the terms excluded as a result of the conceptualization.

The conceptualized terms provide for the establishment of conceptualized vectors. The dimensionality of the vector space obtained was 27. It contained 47 vectors dimensioned from 2 to 27.

**Unsupervised semantic grouping**

The semantic grouping has been performed by applying clustering EM and K-Means algorithms (Alldrin Smity, and Turnbull, 2003) to the conceptualized vector set. Term has been set as both input and predictable column. The results obtained are presented in Table 1.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Minimum support</th>
<th>Cluster count</th>
<th>Clusters with concept vectors</th>
<th>Mixed clusters</th>
<th>Favor topic extraction</th>
<th>Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scalable EM</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>Low</td>
<td>No</td>
</tr>
<tr>
<td>Scalable EM</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>Medium</td>
<td>No</td>
</tr>
<tr>
<td>Scalable K-means</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>Medium</td>
<td>No</td>
</tr>
<tr>
<td>Scalable K-means</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>High</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note: To appear after the paragraph “The semantic grouping...” and before the paragraph “The experiments differ...”
The experiments differ by the minimum support values. With the default value 1 the count of obtained clusters with EM and K-means differ significantly. K-means produces stable cluster count with the two different minimum support values. Both algorithms output one mixed cluster. It contains some conceptual vectors without any conceptual terms.

Both algorithms have been scored for better favoring automatic topic selection. Due to the stable results for cluster count K-means has been scored higher than EM. K-means with the greater support value has been selected for further ontologically supported topic extraction.

**Selection of initial topics by ontological mapping**

The task consists in assigning a topic to each of the established clusters. The algorithm is shown in Figure 3.

![Figure 3. Extraction of topic from a cluster](image)

Each conceptual vector from the cluster is mapped to the ontology taxonomy. Since the vector terms are ontological concepts the algorithm checks whether they match. In case of match the topic is straightforwardly extracted. In case of not match the subsuming concepts from the taxonomy are extracted and a resultant superconcept vector is obtained.

The new concepts are checked for match, and so on until all concepts of the superconcept vector become the same. This derived concept is assigned as topic to the initial concept vector. The procedure is applied to all clusters and in result the conceptual vectors will be labeled with the extracted topics.

Part of the topic vectors obtained by the algorithm is shown in Table 2.
TABLE 2. CONCEPTUAL VECTORS WITH AUTOMATICALLY EXTRACTED TOPICS

<table>
<thead>
<tr>
<th>Vector</th>
<th>Abstract text</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.txt</td>
<td>ANALYTICAL MODELS FOR E-GOVERNANCE Management concerns decision making based on complex analysis of data about organization’s processes flow...</td>
<td>E-Governance</td>
</tr>
<tr>
<td>10.txt</td>
<td>Analysis of the interoperability of information systems IN state ADMINISTRATION Interoperability is a basic element for building e-government. This is a subject...</td>
<td>E-Government structure</td>
</tr>
<tr>
<td>15.txt</td>
<td>Electronic administrative service in the National Revenue Agency - Status and Developments This paper is focused on the achieved results and the possibilities for development of administrative service through ......</td>
<td>E-Government service</td>
</tr>
<tr>
<td>18.txt</td>
<td>FACTORS FOR SUSTAINABILITY OF CENTRES WHICH PARTICIPATE IN EU E-GOVERNANCE NETWORKS The development of e-governance in EU is supported by activities of a lot of e-governance centers...</td>
<td>E-Government service</td>
</tr>
</tbody>
</table>

Model training

The ontologically supervised model has been trained with three machine learning algorithms, i.e. decision tree, Naïve Bayes and neural network with the topic being the predictable attribute. The models obtained by the decision tree and Naïve Bayes algorithms are shown in Figure 4 and Figure 5.
As shown in the figures both classifiers output similar results for the topic predictors.

**Text classification**

The trained models have been used for supervised classification of new texts. The texts have been downloaded from Internet by searching with the keywords e-governance ontology and text classification. A document set with ten abstracts has been prepared as classifier's input. The topic field has been left blank. The set has been conceptualized by the procedure of the described approach. The prediction task consisted in determining the probability that the documents from the set belong to a specified topic. Predictions have been made for the topics “E-governance”, “E-government service” and “E-government structure”. The results obtained by implementing the three ontological classifiers are shown in Figure 6.
All models classified document1 to the topic “E-governance”. No documents were classified to the topic “E-government structure”. Three documents were classified to the topic “E-government service” by the Naïve Bayes classifier only.

Conclusion

Current work represents further extension of the research on a general framework for knowledge generation from e-library the design of which has been initiated in (Tsankova et al., 2011). The approach for performing unsupervised text classification has been developed in (Rozeva et al., 2011). The method of supervised classification of incoming texts to the e-library demanded the proper ontology selection and design of a mechanism for its involvement in the general text classification process. The results obtained so far although on a limited number of the training set are encouraging and prove the applicability of the proposed approach. The models obtained by implementing three different machine learning algorithms for classification show that the e-library documents align sufficiently well to the ontology that has been selected in the supervised model training. The predictions that have been made on a randomly selected document set in our opinion are relevant to their real content. Currently the Naïve Bayes model outperforms other models at the prediction stage.

The approach can be applied for classification of other texts under the supervision of different ontologies. The results shown by now although preliminary are supporting the main research goal. The approach can be further refined at the stages of clustering algorithm selection and topic extraction by mapping concept vectors to the ontology taxonomy.

References


