OmniCat: Automatic Text Classification with Dynamically Defined Categories

Maciej Janik
Large Scale Distributed Information Systems Lab (LSDIS)
Computer Science Department, University of Georgia
Boyd Graduate Studies Research Center, Athens, GA 30602-7404
mjanik@uga.edu

Krys J. Kochut
Large Scale Distributed Information Systems Lab (LSDIS)
Computer Science Department, University of Georgia
Boyd Graduate Studies Research Center, Athens, GA 30602-7404
kochut@cs.uga.edu

ABSTRACT
We present OmniCat, an ontology-based text categorization method that classifies documents into a dynamically defined set of categories specified as contexts in the domain ontology. The method does not require a training set and is based on measuring the semantic similarity of the thematic graph created from a text document and the ontology fragments created by the projection of the defined contexts. The domain ontology together with the defined contexts effectively becomes the classifier, as it includes all of the necessary semantic and structural features of the classification categories. With the proposed approach, we can also dynamically change the classification categories without the need to retrain the classifier. In our experiments, we used an RDF ontology created from the full version of the English language Wikipedia to categorize a set of CNN documents and a subset of the Reuters RCV1 corpora. The high accuracy achieved in our tests demonstrates the effectiveness of the proposed method and applicability of Wikipedia for semantic text categorization.

Keywords
Ontology-based text categorization, Semantic Web

1. INTRODUCTION
Automatic text categorization methods use machine learning or statistical approaches to classify documents to previously defined and learned categories. We, humans, in addition to understanding the document's content, use the general background knowledge about the surrounding world and the interest in certain aspects or categories to perform the categorization task. We recognize named entities, their roles and identify associations between them. In addition to the information contained in the document, we use our knowledge to better understand the document and fill in the contextual facts and relationships among them that are not explicitly stated in the document. Frequently, our interest in certain subjects influences our perceived importance of certain facts in the document. Consequently, we pay more attention to entities and information that are in our context of interest.

We propose to use similar approach for the text classification with the use of the ontology. It directly leverages domain knowledge from the ontology for the task of automatic text categorization and enhance it with the definition of categorization context to capture user's interest. The novelty of our approach is that it does not depend on a set of pre-defined, fixed categories and the associated set of training documents. Instead, it relies exclusively on the knowledge represented in the ontology: (1) named entities, relationships between them, entity classification and the class hierarchy and (2) the dynamically defined ontology contexts, representing the classification categories.

2. CONTEXT - ONTOLOGY SUBGRAPH
Our definition of a categorization context is based on the previous works on the notion of views in semi-structured databases [1] and in ontologies [2]. We define context in terms of an RDF/RDFS ontology.

<u>Def. 1</u>. The *hierarchical distance*, $dist_H(e,c)$, between an instance entity e and a class c is defined as the length of the shortest path formed by one rdf:type and zero or more rdfs:subClassOf properties connecting e and c. If the entity e is not an instance of class c, $dist_H(e,c)$ is set to 0. The hierarchical distance between an instance entity e and a set of classes C is defined as the minimum, positive value of all $dist_H(e,c)$, where $c \in C$. If e is not an instance of any of the classes in C, $dist_H(e,C)$ is set to 0.

<u>Def. 2</u>. Let C be a set of schema classes included in an RDFS schema S. A projection of classes C onto an RDF description base R is a set of instance entities in R together with their associated hierarchical distance to C, defined as:

$$\Pi(C,R) = \{ e(k) : e \in \mathrm{R} \wedge k = dist_H(e,C) \wedge k > 0 \}.$$

<u>Def. 3</u>. The *categorization context* is a projection of a given set of schema classes onto an RDF description base. An instance entity e is covered by a categorization context C, when its hierarchical distance to the context C is greater than zero.

<u>Def. 4</u>. Given two categorization contexts m_1 and m_2, intersection of contexts ($m_1 \cap m_2$), union of contexts ($m_1 \cup m_2$) and the difference of contexts ($m_1 \backslash m_2$) are also categorization contexts

3. CATEGORIZATION ALGORITHM
Our categorization algorithm consists of three main steps described in the outline below. We have modified and extended our previously presented categorization algorithm [3] with dynamically defined classification categories represented by ontology contexts. The flexibility in almost arbitrary specification of the classification contexts allows the user to create different perspectives or views for the categorized document corpora.

Semantic graph construction
- Named entity identification based on phrases describing the entities in the ontology (entity labels), their confidence and strength of the textual match.

- Entity relations extraction between phrases spotted in the analyzed document.
- Connectivity inducement by adding relationships present in the ontology between the spotted entities together with the relationship importance defined in the schema.

Thematic graph selection

- (optional, news specific) Removal of entities of specific types from graph (places, dates or time-related entities).
- Connected component identification in undirected graph.
- Information and weight propagation to establish the most authoritative entities in the graph.
- Dominant thematic graph identification as the largest and most important graph component.
- Selection of the core and most central entities as topic landmarks.

Categorization into defined contexts

- Fitness score calculation to measure the similarity of the thematic graph and the categorization context. It is based on the context coverage (classification category) and its distance from the document's thematic graph is calculated for all categorization contexts.
- Top *n* categories ranked according to their fitness scores are assigned to the document.

3.1 Classification into contexts

Classification of a document into the defined categories (contexts) requires calculating a fitness score of the document's thematic graph for each of the defined contexts. The fitness score of the thematic graph against a given context represents their semantic similarity and is calculated based on the following conditions:

- the intersection of the context projection with the thematic graph is maximized (coverage),
- the hierarchical distance of the entities in the thematic graph to the classes included in the context is minimized (closeness),
- at least one of the core entities is covered by the context (coverage of the core),
- the highest number of the core entities are covered and close (in hierarchical distance) to context.

The fitness score for the thematic graph T and context C is calculated using the following formula:

$$fs(C,T) = \sum_{k} w_k * h(dist_H(e_k, C)) + \sum_{n} w_{cn} * h_c(dist_H(e_{cn}, C))$$

where k is the number of entities and n is the number of core entities in T covered by context C, e_k and e_{cn} are entities and core entities, w_k and w_{cn} are weights of entities e_k and e_{cn}, and the functions h and h_c represent the importance of the entity's distance from context for normal and core entities.

In our experiments, we used normal distribution function N with the mean at 1 and variance 2 as the importance function to favor entities close to the context (distance up to 3) and minimizing the influence of farther entities (with distance 4 and above):

$$h(dist_H(e,C)) = N_{(1,2)}(dist_H(e,C))$$

4. EXPERIMENTS

In our experiments (OmniCat) we used an RDF ontology created from the full version of English Wikipedia XML dump from 08/01/03. We compared the performance of our categorization method with Naïve Bayes from BOW toolkit [5] and SVM from WEKA [6]. We include results from the previous version of the ontology-based categorization that did not use categorization contexts (Onto). We used 2,418 news documents from the CNN (www.cnn.com) RSS feeds (07/07/03–07/09/04) classified into 11 categories, and a subset of 2,254 documents from the Reuters RCV1 [3] corpora (96/08/20-96/09/02) for 6 selected categories. Results of the categorization are presented in Figure 1.

CNN	BOW	SVM	Onto	OmniCat
education	28.5%	28.6%	71.4%	85.7%
health	96.5%	95.4%	54.0%	60.9%
money_autos	46.1%	57.7%	73.1%	80.8%
money_companies	97.5%	96.4%	71.6%	75.6%
money_taxes	41.7%	50.0%	16.7%	33.3%
politics	98.2%	98.8%	88.6%	91.0%
science_and_space	100.0%	100.0%	88.9%	88.9%
sport_mlb	98.8%	95.3%	92.4%	93.0%
sport_nba	95.9%	95.9%	83.6%	86.1%
sport_nfl	100.0%	99.1%	87.8%	89.2%
sport_nhl	97.0%	97.0%	78.0%	80.0%
Total	**95.8%**	**95.4%**	**80.2%**	**83.1%**

REUTERS	BOW	SVM	Onto	OmniCat
Crime & law enforcement	93.1%	95.8%	87.1%	74.3%
Economic performance	88.9%	85.9%	85.9%	83.0%
Elections	96.9%	91.7%	90.1%	92.7%
Health	70.9%	53.2%	59.5%	65.8%
Religion	33.3%	23.5%	58.8%	56.9%
Sports	97.1%	96.6%	92.6%	90.0%
Total	**90.5%**	**88.7%**	**86.1%**	**81.2%**

Figure 1. Comparison of categorization accuracy

5. CONCLUSIONS AND FUTURE WORK

The presented novel approach to text categorization achieved very good results. It relies only on the ontological knowledge and classifier training is not required. The defined categorization contexts allow focusing categorization on specific type of important entities or structures. We intend to extend the definition of categories by allowing linear composition of contexts to capture more complex categorization domains. Also, multiple language versions of Wikipedia open new possibilities for ontology-based categorization, as it relies on entities, relationships and categories, defined as ontological contexts.

6. REFERENCES

[1] Abiteboul, S. et al.: Views for Semistructured Data. Workshop on Management of Semistructured Data, Tucson, AZ, USA (1997)

[2] Decker, S., Sintek, M., Nejdl, W.: The Model-Theoretic Semantics of TRIPLE. Technical Report (2002)

[3] Janik, M., Kochut, K.J.: Wikipedia in Action: Ontological Knowledge in Text Categorization. 2nd Int. Conference on Semantic Computing , Santa Clara, CA, USA (2008)

[4] Lewis, D.D., Yang, Y., Rose, T., Li, F.: RCV1: A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research 5 (2004) 361-369

[5] McCallum, A.K.: Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering.: http://www.cs.cmu.edu/~mccallum/bow (1996)

[6] Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques (2nd ed.). Morgan Kaufmann, San Francisco (2005)