Variable Selection as an Instance-based Ontology Mapping Strategy

Konstantin Todorov[1], **Peter Geibel**[2]
[1]IKW, Universtity of Osnabrück, Albrechstr. 28, 49076 Osnabrück, Germany
[2]TU Berlin, Fakultät IV, Franklinstr. 28/29, 10587 Berlin, Germany

Abstract— *The paper presents a novel instance-based approach to aligning concepts taken from two heterogeneous ontologies populated with text documents. We introduce a concept similarity measure based on the size of the intersection of the sets of variables which are most important for the class separation of the instances in both input ontologies. We suggest a VC dimension variable selection criterion elaborated for Support Vector Machines (SVMs), which is novel in the SVMs literature. The study contains results from experiments on real-world text data, where variables are selected using a discriminant analysis framework and standard feature selection techniques for text categorization.*

1. Introduction

Instance-based or extensional ontology mapping comprises a set of theoretical approaches and tools for measuring the semantic proximity of two ontologies based on their extensions - the instances that populate their concepts. Typically, a set theoretic approach to modeling concepts is adopted: the relatedness of a pair of concepts is an outcome of a properly chosen measure of similarity, usually based on estimations of the intersections of two sets of instances.

There exists already a list of similarity measures to choose from together with mapping systems which employ them. Among the most popular choices is the Jaccard coefficient [4], as well as a couple of standard statistical measures which have been already applied for extracting semantics out of natural texts based on term co-occurrence, such as mutual information, log-likelihood and others [22]. For an overview of instance-based mapping in terms of measures, thresholds and type of concept instantiation[1] we refer to the empirical study carried out by Isaac et al. [8]. The overall topic of ontology matching is covered in the book of the same name by Euzenat and Shvaiko [5].

In the current paper we propose a novel measure of instance-based concept similarity using variable selection for class discrimination. The instances in our study are natural text documents assigned to the nodes of each ontology and coded as TF/IDF vectors [9]. Variable selection mechanisms are used to find variables (terms from the TF/IDF vector dimensions), which are most characteristic for a given concept and play the most important role for separating its instances from the rest of the instances of the same ontology. The proposed measure of similarity is based on comparing the most important variables for two concepts taken from different ontologies. The choice of a variable selection procedure within this setting is left to the user. However, we propose a novel selection criterion elaborated for Support Vector Machines (SVMs), arguing that it potentially outperforms standard selection techniques. The viability of the proposed concept similarity measure is demonstrated by experiments carried out by the help of discriminant analysis (DA) and standard selection techniques for text categorization.

The paper is structured as it follows. The next two sections describe our ontology mapping scenario and review related work. We introduce variable selection and the resulting concept similarity measure in Section 4. Section 5 presents an overview of the SVM classifier, reviews existing SVM-based variable selection procedures and closes with a description of the theoretical and practical grounds of the proposed SVM-based selection method. Finally, an experimental evaluation of the suggested similarity measure is included in Section 6.

2. Ontology Mapping and Concept Similarity

In our study we focus on hierarchical, tree-like ontologies, designed to categorize text documents (web pages) with respect to their contents[2]. We define a hierarchical ontology O as a finite set of concepts C_O and a set of hyponomic (`is_a`) relations holding between these concepts. We use the documents assigned to a given concept as instances of that concept.

The mapping problem in our setting consists in identifying semantic similarities between two heterogeneous input ontologies, each equipped with a set of instances populating their concepts. The proposed approach serves to align pairs of distinct ontology concepts by their degree of semantic proximity, measured on the basis of their extensions by the help of machine learning techniques.

[1]With respect to whether or not inheritance via subsumtion among concepts is taken into account in defining concepts instance-sets, one distinguishes between hierarchical and non-hierarchical instantiation. The former presupposes that concepts inherit the instances of their predecessors in the hierarchy, the latter does not.

[2]The strictly hierarchical structure of the ontology is not relevant to the performance of the proposed measure. It is however important for an overall matching approach developed by the same authors [18].

We proceed to introduce the data sets on which the machine learners will be applied. Let us consider two ontologies O_1 and O_2 together with their corresponding sets of documents $D_1 = \{\mathbf{d}_1^1, ..., \mathbf{d}_{m_1}^1\}$ and $D_2 = \{\mathbf{d}_1^2, ..., \mathbf{d}_{m_2}^2\}$, where each document is represented as an n-dimensional TF/IDF vector[3] and m_1 and m_2 are integers. The documents in both sets D_1 and D_2 are based on the same set of attributes which can be assumed without loss of generality.

Let A be a concept from ontology O_1. We define a training data set $S^A = \{(\mathbf{d}_i^1, y_i^A)\}$, where $\mathbf{d}_i^1 \in \mathbb{R}^n$, $i = 1, ..., m_1$ and y_i^A are labels taking values $+1$ when the corresponding document $\mathbf{d}_i^1$ is assigned to A and -1 otherwise. The labels separate the documents in ontology O_1 into such that belong to the concept A (positive instances) and such that do not (negative instances).

The same representation and training data set can be acquired analogously for any given concept in both input ontologies O_1 and O_2. The similarity between two concepts A and B which belong to two different ontology will be assessed by the help of their corresponding datasets S^A and S^B.

3. Related Work

We review a couple of related approaches. FCA-MERGE, based on Formal Concept Analysis was proposed by *Stumme and Mädche* [17]. The approach relies on the assumption that two ontologies use the same instances taken from a set of text documents relevant to both of them. It provides its own mechanisms of extracting instances from text corpora, answering a basic critique that source ontologies are unlikely to share the same sets of instances. The approach applies natural language processing and FCA to derive a concept lattice which is further transformed into a merged ontology.

A couple of state-of-the-art solutions are based on machine learning techniques. The instance-based ontology mapper GLUE, introduced by *Doan* and co-workers, utilizes machine learning techniques for deriving semi-automatically assertions on the concepts' similarity [4]. *Lacher and Groh* [10] contributed to the instance-based research by their system CAIMAN which was created to facilitate the retrieval and publishing of documents among communities. An interesting recent approach proposed by *Wang* and colleagues [20] consists in replacing the mapping problem by a classification one by introducing a *similarity space* in which every point represents a pair of matched concepts. Assigning to correct and incorrect matches respectively positive and negative labels allows for the automatic classification of new pairs of concepts as either similar or dissimilar.

We will cite two contributions relying on structure-based techniques. The ANCHOR-PROMPT algorithm developed by *Noy* and colleagues [14] uses a standard graph representation of ontologies. The algorithm starts by selecting a set of pairs of similar concepts from both ontologies - the so called "Anchors" - usually identified through lexical matching. The procedure further builds on the idea that if there have been found two pairs of similar concepts and there exists a path connecting the concepts in each of the two ontologies, it is very likely that the entities found on those paths are also similar. *Mitra and Wiederhold* [13] introduced formally the ontology-composition algebra within the ONION tool for ontology articulation. The authors argued against the need and possibility of constructing and maintaining a global consistent ontology, instead they suggested mechanisms for locally merging parts of ontologies for the purposes of a given application.

A procedure combining instance-based and structural similarity measures was introduced in [18] by the authors of the current paper.

We sum up the advantages and differences of our approach compared to the ones discussed above. The fact that the approach is entirely accomplished at the test phase[4] of the learning task is a serious advantage of the method compared to state-of-the-art approaches (e.g. [4]). In contrast to most instance-based mapping techniques, the presented approach does not rely on instance sets intersection. In fact, it works with document sets that might be different for both ontologies (as seen in the preliminary experiments) which makes the expensive step of extracting instances for the source ontologies from text (as done in [17]) unnecessary. The relevant variables are determined for each ontology independently, and the matching itself is an inexpensive computation. Finally, the method is stable in multi-linguistic environments since documents from both ontologies need not be in the same natural language. It suffices that the documents TF/IDF vectors are translated into a single target language and not even all their features, but only the selected ones.

4. Concept Similarity via Variable Selection

Variable, or *feature selection* is a core problem in a number of real life statistical analysis problems, particularly classification tasks. The result of a variable selection procedure is a list of the input variables, ordered by importance (or *informativeness*) for the output variables (in classification these are the class labels in a training dataset), according to a certain evaluation criterion. On the one hand, this procedure leads to reducing the dimension of the input space ensuring better computational efficiency and improving generalization. On the other hand, in various domains of application,

[3]Alternative representations, such as raw counts of term occurrences and term frequencies, have been used in the experimental studies, as well.

[4]An automatic classification task is typically accomplished in two main steps: *test (or training) phase*, when available data is "learned" by the machine algorithm and *classification phase* when the learned rule is applied on unseen instances.

such as text categorization, process control, gene selection and other, it is important to find out more about the input - output relation in a given data set by pointing out the input variables, which most strongly affect the response. The focus in our study falls on the latter application of variable selection. In that scenario, for a given data set of the type S^A variable selection would indicate which of the TF/IDF vector dimensions are most important for the separation of the documents into such that belong to the concept A and such that do not.

For an overview of general variable selection applications and existing theoretical approaches we refer to the study of Guyon and Elisseeff [6]. Variable selection methods for text-learning have been discussed and evaluated in [12]. SVM-based methods, which are directly relevant to our approach, are discussed separately in Section 5.2 of this paper.

We are coming to the core of the paper: how can a variable selection procedure be applied to discovering *concept similarities*. We take as an input two concepts $A \in C_{O_1}$ and $B \in C_{O_2}$ together with their corresponding datasets $S^A = \{(\mathbf{d}_i^1, y_i^A)\}$, $i = 1, ..., m_1$ and $S^B = \{(\mathbf{d}_j^2, y_j^B)\}$, $j = 1, ..., m_2$ as introduced in Section 2. Our goal is to identify the degree of similarity between these two concepts. We carry out a variable selection procedure on each of the sets and order the variables by their importance for the class separation. Let

$$L^A = \{Var_{\sigma(1)}, Var_{\sigma(2)}, ..., Var_{\sigma(n)}\}$$

and

$$L^B = \{Var_{\delta(1)}, Var_{\delta(2)}, ..., Var_{\delta(n)}\}$$

be the ordered lists of variables for concepts A and B, respectively, where σ and δ are two permutations on the sets of variable indexes. We take from each of the lists a subset of the top k elements, where $k < n$ is to be set by the user and define the subsets $L_k^A = \{Var_{\sigma(i_1)}, Var_{\sigma(i_2)}, ..., Var_{\sigma(i_k)}\}$ and $L_k^B = \{Var_{\delta(j_1)}, Var_{\delta(j_2)}, ..., Var_{\delta(j_k)}\}$, $i, j \in (1, n)$ (Figure 1). The similarity of concepts A and B is given as

$$sim(A, B) = \frac{|L_k^A \cap L_k^B|}{k}, \tag{1}$$

with $sim(A, B) \in (0, 1)$.

The concept or its complement?

Due to the nature of the introduced concept similarity criterion, there appears a certain ambiguity in the final similarity judgment. If a subset of variables is important for the separation of a given data set into classes B and $\overline{B}$ so is the same subset when we swap the two labels. The end result is that whenever our similarity measure $sim(A, B)$ yields 1 or a number close to 1 the following disjunction holds: "concept A is similar to concept B" or "concept A is similar to concept $\overline{B}$" (the second possible disjunction, namely "$\overline{A}$ similar to $\overline{B}$" or "$\overline{A}$ similar to B" is complementary to the first one). We suggest to address this problem by the help of an approach which computes the statistical correlation between an attribute and the corresponding binary output estimated over the training data in order to get the desired sign information.

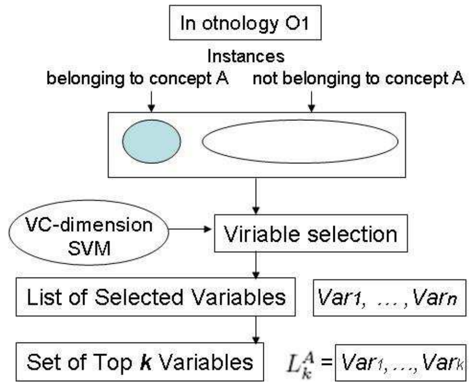

Fig. 1: Variable selection for a concept A in ontology O_1.

5. VC-dimension-based Variable Selection for SVMs

The similarity measure (1) makes use of a variable selection procedure of some kind. In this section we introduce a novel selection criterion based on variations of the VC dimension of a support vector machine classifier. We start by a brief overview of the learning technique.

5.1 Overview of Support Vector Machines

The Support Vector Machines are supervised learning classification techniques introduced in the mid 1990s by Vapnik and coworkers [19]. For reasons of space, we cannot give detailed account of all aspects of SVMs, which combine results from several mathematical fields. Instead, we will provide enough knowledge about the method in order to understand the ideas behind SVM-based variable selection approaches developed in the past decade, as well as to be able to introduce our method. For a thorough overview of SVMs we refer to the book by Cristianini et al. [3].

Let us consider the following binary classification layout. Assume we have l observations $\mathbf{x}_i \in \mathbb{R}^n$ and their associated "truth" $y_i \in \{-1, 1\}$. Data are assumed to be i.i.d. (independent and identically distributed), drawn from an unknown probability distribution $P(\mathbf{x}, y)$. The goal of binary classification is to "learn" the mapping $\mathbf{x}_i \rightarrow y_i$ which is consistent with the given examples. Let $\{f(\mathbf{x}, \alpha)\}$ be a set of such possible mappings, where α denotes a set of parameters. Such a mapping is called a classifier and it is deterministic - for a certain choice of $\mathbf{x}$ and α it will always give the same output f.

The **actual risk**, or the expectation of the test error for such a learning machine is

$$R(\alpha) = \int \frac{1}{2}|y - f(\mathbf{x}, \alpha)|dP(\mathbf{x}, y).$$

The quantity $1/2|y - f(\mathbf{x}, \alpha)|$ is called *the loss*. Based on a finite number of observations, we calculate the **empirical risk**

$$R_{emp}(\alpha) = \frac{1}{2l}\sum_{i=1}^{l}|y_i - f(\mathbf{x}_i, \alpha)|,$$

which is a fixed number for a given training set $\{\mathbf{x}_i, y_i\}$ and a certain choice of parameters α.

For losses taking values 0 or 1, with probability $1 - \eta$, $0 \leq \eta \leq 1$, the two risks are related in the following manner:

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h\log(\frac{2l}{h}) + 1 - \log(\frac{\eta}{4})}{l}}, \quad (2)$$

where h is a nonnegative integer which will play a core role in our variable selection procedure, called the *VC dimension*. The bound (2) gives an insight on one very important aspect of generalization theory of statistical learning. The term $\sqrt{\frac{hlog(\frac{2l}{h})+1-log(\frac{\eta}{4})}{l}}$, called *VC confidence* is "responsible" for the *capacity* of the learner, i.e. its ability to learn unseen data without error. The other right-hand quantity in (2) - the empirical risk, measures the *accuracy* attained on the particular training set $\{\mathbf{x}_i, y_i\}$. What is sought for is a function which minimizes the bound on the actual risk and thus provides a good balance between capacity and accuracy - a problem known in the literature as *capacity control*.

The presented risk bound does not depend on $P(\mathbf{x}, y)$ and it can be easily computed provided the knowledge of h. We introduce what does this parameter stand for. Let us consider the set of functions $\{f(\mathbf{x}, \alpha)\}$ with $f(\mathbf{x}, \alpha) \in \{-1, 1\}, \forall \mathbf{x}, \alpha$. In a binary classification task there are 2^l possible ways of labeling a set of l points. If for each labeling there can be found a member of $\{f(\alpha)\}$ which correctly assigns these labels, we say that the given set of points is *shattered* by the given set of functions. The VC dimension is a property of such a family of functions, which is defined as the maximum number of training points that can be shattered by that family.

We come back to binary classification with support vector machines. SVMs are based on a family of linear functions $\{f(\mathbf{x}, \alpha)\}$ mapping elements from the input space to a binary output, as introduced so far with α being the parameters of the linear function $f(\mathbf{x})$. The classification decision is according to the sign of the linear function at the point to be mapped. Geometrically, it can be thought of as a hyperplane separating the space of the inputs in two halves in a way that the margin between the two classes is maximized.

More formally, let us consider the input space $X \subseteq \mathbb{R}^n$ and the output domain $Y = \{-1, 1\}$ with a training set $S = ((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_l, y_l)) \in (X, Y)^l$. SVM is a linear real function $f : X \rightarrow \mathbb{R}$ with

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b,$$

where $\alpha = (\mathbf{w}, b) \in \mathbb{R}^n \times \mathbb{R}$. The separating hyperplane in the input space X is defined by the set $\{\mathbf{x}|f(\mathbf{x}) = 0\}$. The decision rule assigns an input vector $\mathbf{x}$ positive if and only if $f(\mathbf{x}) \geq 0$ and negative - otherwise. (The inclusion of 0 in the first case and not in the second is conventional.)

We are looking for the best decision function $f(\mathbf{x})$ which separates the input space and maximizes the distance between the positive and negative examples closest to the hyperplane. The parameters of the desired function are found by solving the following quadratic optimization problem:

$$\min_{\mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2$$

under the linear constraints

$$\forall i = 1, ..., n, \; y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle) + b) \geq 1.$$

When data are not linearly separable in the input space, they are mapped into a (possibly higher dimensional) space, called *feature space* where a linear boundary between both classes can be found. The mapping is done implicitly by the help of a kernel function which plays the role of a dot product in the feature space.

5.2 A VC-dimension-based Selection Criterion

The SVMs have many attractive sides - their performance does not depend on the distribution of the data (safe that they are i.i.d.), it does not demand a linear input-output relation and they are easy to implement. At least theoretically, the generalization properties of SVMs do not dependent on the size of the input space which makes variable selection little prominent for learning with SVMs. However, the listed properties turn them into a good candidate for a variable selection tool to be used self-dependently. In addition to that, some authors have shown that even though theoretically unnecessary, variable selection improves SVM learning in practice [15].

SVM-based variable selection has already been studied in the past couple of years. In 2000 Guyon *et al.* proposed the SVM-RFE algorithm [7] for selecting genes which are relevant for cancer classification. The removal criterion for a given variable is minimizing the variation of the weight vector $\| \mathbf{w} \|^2$, i.e. its sensitivity with respect to a variable. Rakotomamonjy *et al.* (2004) carried out experiments for pedestrian recognition by the help of a variable selection procedure for SVMs based on the sensitivity of the margin according to a variable. The guiding idea of their approach is: *"A variable which is little informative and thus little important for the decision function, is a variable to which the margin* $2/\| \mathbf{w} \|$ *is little sensitive."* [15], [16]. A method based on finding the variables which minimize bounds on

kernel	dot	radial
Data:	VCdim <= 15.178417	VCdim <= 1383.9332
Data(01-04):	VCdim <= 17.519558	VCdim <= 1383.8331
Data(05-08):	VCdim <= 27.201857	VCdim <= 1383.2865
Data(09-12):	VCdim <= 17.714419	VCdim <= 1383.5678
Data(13-16):	VCdim <= 16.391693	VCdim <= 1383.9423
Data(17-20):	VCdim <= 16.881794	VCdim <= 1383.9564
Data(21-23):	VCdim <= 11.398824	VCdim <= 1383.5026

Fig. 2: Various VC dimensions estimated over a partitioned data set with two different kernels.

the *leave-one-out* error for classification was introduced by Weston *et al.* in 2000 [21]. Bi *et al.* (2003) developed the VS-SSVM variable selection method for regression tasks applied to molecules bio-activity prediction problems [2].

The variable selection criterion that we propose is based on the sensitivity of the VC dimension of the SVM classifiers with respect to a single variable or a block of variables. As we have seen in Section 2, for different values of the VC dimension h, different values of the VC confidence (describing the capacity of the classifier) will be computed and thus different bounds on the actual risk (2), where from the generalization power of the classifier will change. Our main heuristics can be formulated as *"a less informative variable is one, which the VC confidence of the classifier is less sensitive to"*.

For computational reasons the evaluation function of our variable selection procedure will be formulated in terms of VC dimension directly, instead of in terms of the VC confidence. This is plausible since the VC confidence is monotonous in h. Thus, the i-th variable is evaluated by

$$eval_i = |h(H) - h(H^{(i)})|,\ i = 1, ...n, \tag{3}$$

where $h(H)$ is the VC dimension of a set of SVM hypotheses H constructed over the entire data set and $h(H^{(i)})$ is the same quantity computed after the removal of the i-th variable in the data set (this is the variable whose pertinence is to be evaluated).

We have run experiments in support of the presented evaluation function in the domain of advanced process control. We made observations over production items going through a manufacturing line. A set of variables is assigned to each item during the production process - measurements taken at different points of the process. At the end of the line a certain part of the products have been classified as "defect" (failed to meet the quality requirements) and the rest - as "good". The task was to identify which are those variables the variation of which has caused that some of the items failed to turn out "good". We trained a SVM over the data consisting of input observations and a final binary output ("defect" or "good"). The dataset considered here consists of 23 real variables observed over more than 1000 examples. The training process was repeated 6 times, consequently removing a block of 4 or 3 variables at a time. The blocks have been selected randomly. The corresponding estimations of the VC dimension[5] at each training phase have been measured and then compared to the estimated VC dimension of the whole data set (with all 23 variables included). The values of the observed VC dimensions are given on figure 2 where the variables which have been excluded from the data on each step are in brackets. After applying the ranking criterion introduced in (3) we concluded that the most important variables are contained in the block (05-08). A similar selection procedure has been carried afterwards by consequently removing each of the four variables of that block in order to find out the most significant one(s) among them. The achieved results were in correspondence with the intuitive guesses of the process control engineers.

6. Experiments

While in the process of implementing the SVM approach, in order to demonstrate the viability of the proposed variable-selection-based concept similarity measure (1), we carried out experiments by the help of a couple of standard variable selection techniques.

Experiment 1.

We started by testing the variables importance by carrying out a discriminant analyses (DA). DA is a basic data analysis method which reveals important structural information contained in the data. It is based on constructing principle axes, which capture the separation of the classes by minimizing their in-class variation and maximizing the distances between their means. The resulting principle (discriminant) axes are linear combinations of the input variables, where the variables with greatest weights for the construction of a given axis are most important for the class separation projected on this axis. Therefore, DA analysis can serve as a variable selection tool in class discrimination problems.

We used data from the publicly available "20 Newsgroups" dataset [1] which is a collection of approximately 20,000 news articles, partitioned evenly across 20 different topics. We started with the topics "Autos" and "Religion" and split the documents in "Autos" in two - `Autos1` and `Autos2`, producing sets of instances of two similar pseudo concepts. The documents in `Religion` were used as instances from a third (dissimilar) concept. Our goal was to show that the features which are important for the separation of `Autos1` and `Religion` are the same as those important for the separation of `Autos2` and `Religion`. We carried out a DA on the data set consisting of the three categories of TF/IDF documents, introduced so far. Figure 3 shows the results of the analysis on the first two discriminant axes. (The labels on the plot are as it follows: (1) for `Autos1`, (2) for `Autos2` and (-1) for `Religion`.) The

[5] In general, it is difficult to compute the VC-dimension directly, but in the case of SVMs, we can compute an upper bound for it depending on the resulting weight vector and on properties of the given data. In the experiments, we used that upper bound.

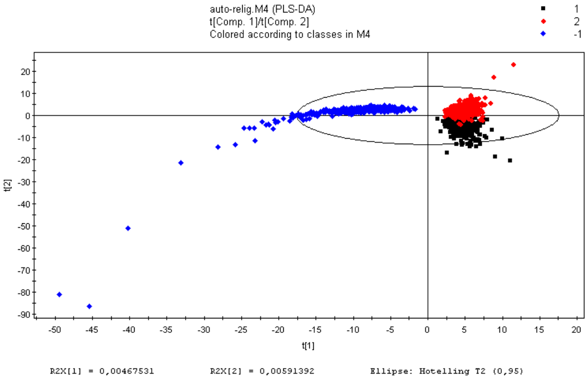

Fig. 3: A DA projection of the population of documents from three classes onto the first two discriminant axes.

Auto1 vs. Rel1+Pol1			Auto2 vs. Rel2+Pol2			Rel2 vs. Auto2+Pol2			Pol2 vs. Auto2+Rel2		
	1	2		1	2		1	2		1	2
1	Var ID	M1.VIP[6]	1	Var ID	M1.VIP[5]	1	Var ID	M1.VIP[4]	1	Var ID	M1.VIP[5]
2	Var_13	12,0871	2	Var_13	13,5142	2	Var_4239	16,401	2	Var_4823	13,7799
3	Var_1712	8,50679	3	Var_1712	8,64681	3	Var_6766	16,3152	3	Var_4822	12,9308
4	Var_4239	8,49854	4	Var_4239	8,57421	4	Var_4470	16,3066	4	Var_1002	12,0012
5	Var_6767	8,47135	5	Var_6766	8,54044	5	Var_6767	16,2867	5	Var_4239	7,877
6	Var_6766	8,46965	6	Var_4470	8,53651	6	Var_1712	15,9402	6	Var_6766	7,82062
7	Var_4470	8,41407	7	Var_6767	8,52678	7	Var_4428	15,7792	7	Var_4470	7,81634
8	Var_104	8,36099	8	Var_4428	8,27657	8	Var_4443	12,8257	8	Var_6767	7,80659
9	Var_4428	8,34646	9	Var_1002	8,07467	9	Var_6763	11,7615	9	Var_4428	7,63906
10	Var_4823	8,00595	10	Var_4823	7,94405	10	Var_6771	10,307	10	Var_1712	7,38678
11	Var_4443	8,00247	11	Var_4443	7,71282	11	Var_155	10,127	11	Var_5191	6,50381
12	Var_4822	7,74572	12	Var_4822	7,66364	12	Var_20	8,89759	12	Var_13	6,09668
13	Var_1002	7,72475	13	Var_104	7,05156	13	Var_148	8,5031	13	Var_220	5,94339
14	Var_6771	6,72282	14	Var_118	6,6756	14	Var_288	8,27832	14	Var_155	5,7131
15	Var_677	5,59383	15	Var_6763	6,14793	15	Var_4823	7,44476	15	Var_11352	5,6608
16	Var_222	5,3653	16	Var_370	5,89779	16	Var_3	6,85242	16	Var_4748	5,64234
17	Var_118	5,18424	17	Var_6771	5,39264	17	Var_4822	6,73377	17	Var_6763	5,62866
18	Var_209	4,83134	18	Var_209	5,23991	18	Var_13	6,57272	18	Var_3763	5,59995
19	Var_6763	4,81526	19	Var_630	5,22241	19	Var_2437	6,51098	19	Var_511	5,53548
20	Var_288	4,69759	20	Var_222	4,58835	20	Var_2974	6,38466	20	Var_20	5,50192
21	Var_217	4,64516	21	Var_102	4,41872	21	Var_143	6,31536	21	Var_1991	5,38238
22	Var_148	4,59836	22	Var_155	4,40869	22	Var_6776	6,15808	22	Var_4443	5,33161
23	Var_102	4,58654	23	Var_496	4,34491	23	Var_90	6,05312	23	Var_1111	5,04141

Fig. 4: The top 23 characteristic variables for four concepts in four different DAs.

two Autos classes appear very close to one another, sharing a big overlap and clearly separated from the Religion class. This observation shows that DA is not able to discriminate properly between the two Auto classes, but separates them well from the Religion class, i.e. the same separation criteria hold for the classes `Autos1` and `Religion` as for the classes `Autos2` and `Religion`.

Experiment 2.

To reinforce this finding, we took a third class from the 20 Newsgroups - "Politics" and split its instances in two, producing two similar concepts out of it. The same was done with the instances in Religion. We mimicked two ontologies, each containing three concepts: $O_1 = \{\texttt{Autos1}, \texttt{Religion1}, \texttt{Politics1}\}$ and $O_2 = \{\texttt{Autos2}, \texttt{Religion2}, \texttt{Politics2}\}$. Let us recall our main argumentation: for separating similar classes we need similar attributes, while for separating dissimilar classes we need a dissimilar set of attributes. Our goal was to evaluate the similarity of concepts `Autos1` and `Autos2` and the dissimilarity of concepts `Autos1` and `Politcs2` and `Autos1` and `Religion2` by applying the measure (1). To that end, we carried out a DA and selected the important variables for the class separation in four analyses:

(DA1) `Autos1` vs. (`Religion1` + `Politics1`) - find the important variables that separate `Autos1` from all other concepts in O_1;

(DA2) `Autos2` vs. (`Religion2` + `Politics2`) - find the important variables that separate `Autos2` from all other concepts in O_2;

(DA3) `Religion2` vs. (`Autos2` + `Politics2`) - find the important variables that separate `Religion2` (a dissimilar concept) from all other concepts in O_2;

(DA4) `Politics2` vs. (`Autos2` + `Religion2`) - find the important variables that separate `Politics2` (a dissimilar concept) from all other concepts in O_2;

Figure 4 shows the lists of the top 23 most important variables for the class separation in the four different DA analyses. (VIP stands for a score coefficient calculated on the basis of the contribution of a single variable to the construction of the discriminant axes.) The result is that the lists of variables separating the classes in analyses (DA1) and (DA2) are very similar, almost identical, where as the variables separating the dissimilar concepts in analyses (DA3) and (DA4) differ from the lists obtained in the first two analysis. (We note that they do share a small overlap, for the concepts are not totally dissimilar, but rather.) By applying our variable-selection-based measure of similarity, we conclude that the concept `Autos1` from O_1 is similar to the concept `Autos2` from O_2 and dissimilar to the concepts `Religion2` and `Politics2` from O_2 which is in line with the semantical nature of the selected classes.

Table 1: Performance using Mutual Information

Concept Names	HW:Mixed	Autos	Religion2	Politics2
HW:PC	**0,033**	0	0	0
HW:Mac	**0,067**	0	0	0
Religion1	0	0	**0,3**	0
Politics1	0	0	0	**0,3**

Experiment 3.

Finally, we have carried an additional study by the help of three other standard variable selection techniques: Mutual Information, Chi-square statistics and Document Frequency Thresholding. The three methods are described in [22], for space limitations we will not discuss them here. By using the "20 Newsgroups" dataset again we mimicked the following two ontologies (the abbreviation "HW" stands for "Hardware"): $O_1 = \{\texttt{HW:PC}, \texttt{HW:Mac}, \texttt{Religion1}, \texttt{Politics1}\}$ and

Table 2: Performance using Chi^2

Concept Names	HW:Mixed	Autos	Religion2	Politics2
HW:PC	**0,700**	0,433	0,400	0,367
HW:Mac	**0,500**	0,467	0,433	0,367
Religion1	0,400	0,400	**0,700**	0,300
Politics1	0,333	0,367	0,333	**0,633**

Table 3: Performance using DF Thresholding

Concept Names	HW:Mixed	Autos	Religion2	Politics2
HW:PC	**0,722**	0,556	0,440	0,485
HW:Mac	**0,726**	0,545	0,437	0,489
Religion1	0,431	0,444	**0,753**	0,541
Politics1	0,479	0,526	0,550	**0,772**

$O_2 = \{\texttt{HW:Mixed}, \texttt{Autos}, \texttt{Religion2}, \texttt{Politics2}\}$. We have chosen the concepts and the documents for our task in such a manner that there are pairs of concepts which are clearly similar (e.g. `Religion1` and `Religion2`) and pairs of concepts which are clearly dissimilar (e.g. `Religion1` and `Autos`). In addition, there is one concept from O_2 which is in a way the union of two concepts of O_1 (the concepts `HW:Mixed` and the concepts `HW:PC` and `HW:Mac`). Each of the classes contains approximately 500 distinct documents on the corresponding topic, none of the classes contains documents that are contained in another class. The results of applying the similarity measure (1) are shown on Tables 1, 2 and 3 in three similarity matrices (one for each variable selection technique applied). The results clearly show that in all three cases a greater similarity is attributed to the concept pairs which are heuristically expected to be more similar, as compared to the expectedly dissimilar concepts.

7. Conclusion and Future Work

The paper presents an instance-based approach to aligning concepts taken from two heterogeneous ontologies populated with documents. It introduces a concept similarity measure based on the class separation information in both input ontologies provided by selecting most important variables. We propose a VC-dimension-based variable selection procedure for SVMs in order to extract the desired information from the instances populating the two input ontologies.

The introduced similarity measure can be successfully applied by the help of any appropriate variable selection procedure instead of the proposed one, as this is seen from our experiments. However, a task of future work is implementing the SVM-based approach, since working with SVMs has many benefits, which have been pointed out in Section 5.2.

References

[1] http://people.csail.mit.edu/jrennie/20Newsgroups/

[2] J. Bi, K. Bennett, M. Embrechts, C. Breneman, M. Song. Dimensionality reduction via sparse support vector machines, *Journal of Machine Learning Research*, 1229–1243, volume 3, MIT Press, 2003.

[3] N. Cristianini, J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods.*, Cambridge University Press, ISBN 0-521-78019-5, 2000.

[4] A. Doan, J. Madhavan, P. Domingos, A. Halevy. Learning to map between ontologies on the semantic web, In *The Eleventh International WWW Conference*, Hawaii, US, 2002.

[5] J. Euzenat, P. Shvaiko. *Ontology Matching*, Springer-Verlag New York, Inc., 2007.

[6] I. Guyon, A. Elisseeff. An introduction to variable and feature selection, *J. Mach. Learn. Res.*, vol. 3, 1157–1182, MIT Press, 2003.

[7] I. Guyon, J. Weston, S. Barnhill, V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines, *J. Mach. Learn.*, vol. 46, 389-422, Kluwer Academic Publishers, 2002.

[8] A. Isaac, L. van der Meij, S. Schlobach, S. Wang. An empirical study of instance-based ontology matching. In *Proceedings of the 6th International Semantic Web Conference*, Busan, Korea, 2007.

[9] T. Joachims. Text categorization with support vector machines: learning with many relevant features. *Proceedings of ECML-98, 10th European Conference on Machine Learning*, Number 1398, 137-142, 1998.

[10] M. Lacher, G. Groh. Facilitating the exchange of explicit knowledge through ontology mappings, In *Proceedings of the 1,ith International FLAIRS conference*, Key West, FL, USA, May 2001.

[11] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, K.R. Mullers. Fisher discriminant analysis with kernels, *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 41-48, 1999.

[12] D. Mladenic. Feature subset selection in text-learning, *Machine Learning: ECML-98*, Volume 1398/1998, Pages 95-100, 1998.

[13] P. Mitra, G. Wiederhold, M. Kersten. A Graph-Oriented Model for Articulation of Ontology Interdependencies, *Lecture Notes in Computer Science*, vol. 1777, p. 86+, 2000.

[14] N. Noy, M. Musen. Anchor-prompt: Using non-local context for semantic matching. In *Proc. IJCAI 2001 workshop on ontology and information sharing*, pages 63Ű70, 2001.

[15] A. Rakotomamonjy. Variable selection using svm based criteria, *J. Mach. Learn. Res.*, vol. 3, 1357–1370, MIT Press, 2003.

[16] A. Rakotomamonjy, F. Suard, Selection de variables par SVM: application a la detection de pietons, *RFIA04*, 2004.

[17] G. Stumme, A. Maedche. FCA-MERGE: Bottom-Up Merging of Ontologies, *IJCAI*, 225-234, 2001.

[18] K. Todorov, P. Geibel. Ontology Mapping via Structural and Instance-Based Similarity Measures. In *4th International Ontology Matching Workshop, 7th International Conference on the Semantic Web*, Karlsruhe, 2008.

[19] V. N. Vapnik. *The Nature of Statistical Learning Theory (Information Science and Statistics).* Springer, New York, 1999.

[20] S. Wang, G. Englebienne, S. Schlobach. Learning concept mappings from instant similarity, In *Proceedings of the 7th International Semantic Web Conference*, pages 339-355, Karlsruhe, 2008.

[21] J. Weston, S. Mukherjee, O.Chapelle, M. Pontil, T. Poggio, V. Vapnik. Feature selection for SVMs, In *Neural Information Processing Systems*, volume 13, p. 668-674, 2000.

[22] Y. Yang, J. O. Pedersen. A comparative study on feature selection in text categorization, In *International Conference on Machine Learning*, pages 412-420, 1997.