Detecting Ontology Mappings via Descriptive Statistical Methods

Konstantin Todorov
Institute of Cognitive Science
Universtity of Osnabrück, Albrechstr. 28, 49076 Osnabrück, Germany
ktodorov@uos.de

Abstract

Instance-based ontology mapping comprises a collection of theoretical approaches and applications for identifying the implicit semantic similarities between two ontologies on the basis of the instances that populate their concepts. The current paper situates this general problem in the realm of finding mappings between the nodes of two different web-directories populated with text documents (the web pages that they intend to organize). We propose a novel approach to detect potential concept mappings based on Principle Component Analysis and Discriminant Analysis and introduce a resulting concept similarity measure. The procedure can be used as an independent concept mapping technique, or as a support to a concept similarity measure of other nature.

1. Introduction

Ontologies describe the semantics of data in order to provide a uniform framework of understanding between different parties. The main common reference to an ontology definition was provided by Gruber in 1993, describing ontologies as knowledge bodies which bring a formal representation of a shared conceptualization of a domain - the objects, concepts and other entities that are assumed to exist in a certain area of interest together with the relationships holding among them [8]. The core-bodies of ontologies are taxonomies - hierarchical structures that organize concepts by subsumptional (is-a) relations. Web directories, such as Yahoo or the Open Directory Project are examples of taxonomies which classify items in a given domain of interest. In the sequel we will speak of (hierarchical) ontologies, but let us keep in mind that the intended application of our approaches is handling web directories formalized as hierarchical ontologies where each ontology concept corresponds to a directory category.

The problem of ontology mapping is rooted in the fact that the nature of ontology acquisition is decentralized and strongly human-biased. This has lead to the creation of a considerable number of ontologies which describe similar or overlapping domains of knowledge but their elements do not explicitly match. Ontology matching can be defined as identifying the implicitly contained similarities between the elements of two heterogeneous ontologies.

In the current paper we will focus on one general type of ontology mapping - instance based mapping. Known also as extensional mapping, it comprises a set of approaches for measuring the semantic similarity of two ontologies based on their extensions - the instances that populate their concepts. Commonly, a set theoretic approach to modeling concepts is adopted in which the relatedness of a pair of concepts is estimated on the basis of the intersection of their instance sets.

How is the set of instances of a given concept defined? Assuming that we have a set of annotated instances for each ontology (i.e. for each instance there is a pointer to which ontology concept it refers to, doubly annotated instances are considered separately and independently for each ontology), there are two possibilities. The first one is to ignore the hierarchical structure of the underlying taxonomy and take as instances of a given concept only those that are directly assigned to it and let us call that a *non-hierarchical instantiation*. The second possibility is to include in the set of instances of a concept all instances assigned to that concept and all of its descendants in the concept hierarchy - a *hierarchical instantiation* [9]. Usually, the choice of one of the two types of instantiations is motivated by semantic considerations dependent on the particular application.

After we have a definition of a set of instances of an ontology concept, an instance-based similarity measure can be introduced in order to yield assertions on the intentional closeness of two concepts taken from two different ontologies. In the current paper we present a novel technique for identifying potential mappings between concepts, based on Principle Component Analysis and Discriminant Analysis. It relies on discovering similarities according to the structure of both input ontology instance sets. The procedure can be used self-dependently, or in combination with another mapping technique. In the second case it serves as a procedure for narrowing down the number of concepts considered as candidates for a semantic similarity check (performed by a certain concept similarity measure). Finally, we propose a concept similarity measure, which comes as a natural output of the analytical methods, based on selecting and comparing the most important variables separating the concepts within the ontologies.

The paper is organized in the following manner. Section 2 provides some background knowledge and assumptions of our model. Related work is reviewed in Section 3. Our novel approach to concept similarity detection and the resulting similarity measure are discussed in Section 4, followed by the results of an empirical study (Section 5).

2 The ontology mapping scenario

In order to discuss formally the problem of instance-based directory mapping, let us start by introducing the definition of a directory (a hierarchical ontology) that we will adopt.

Definition 1 *A hierarchical ontology is a pair* $O := (C_O, \texttt{is\_a})$*, where* C_O *is a finite set whose elements are called concepts and* $\texttt{is\_a}$ *is a partial order on* C_O *with the following property:*

- there exists exactly one element $A_0 \in C_O$ *such that* $\{A_1 \in C_O | (A_0, A_1) \in \texttt{is\_a}\} = \emptyset$*,*

- for every element $A \in C_O$*,* $A \neq A_0$*, there exists an unique element* $A' \in C_O$ *such that* $(A, A') \in \texttt{is\_a}$.

Note that the given definition is purely intentional and does not imply the existence of instances of the ontology concepts. So let us make it explicit that we assume the existence of an extension of each ontology which is a collection of annotated text documents (web-pages) assigned to that ontology and distributed among its nodes. For an ontology O, let its corresponding document set be $D_O = \{\mathbf{d}_1^O, ..., \mathbf{d}_{m_O}^O\}$, where each element $\mathbf{d}_i^O$, $i = 1, .., m_O$ is a text document represented as an n-dimensional TF/IDF feature vector as described by Joachims in [10] and m_O is an integer.

The problem of ontology mapping can be formalized in the following framework (slightly modifying [6] and [9]). Let O_1 and O_2 be two ontologies. For a concept $A \in O_1$ and a concept $B \in O_2$, a mapping is defined as the triple $M_{AB} = M(A, B, R)$, where R is a relation holding between the two concepts ranging from ”identical” ($\equiv$) to ”disjoint” ($\perp$).

3 Related Work

In many open and evolving systems and applications, such as Peer-2-Peer Systems, eCommerce or the widely discussed Semantic Web [3], it has become an urgent task to develop approaches to reconcile heterogeneous ontologies in order to unlock the potential and fully enable the functionality of these systems. Researchers and practitioners have tried to find solution to this problem and a number of theoretical and practical approaches are already out there. The recent ontology matching book by Euzenat and Shvaiko [6] and the survey by Kalfoglou and co-workers [12] are useful references to the topic.

A couple of prominent ontology merging attempts include Noy’s Protégé Prompt tool [17] and the FCA-Merge approach by Stumme and Mädche [18], based on extracting formal contexts out of natural text documents collections. Mitra and Wiederhold [16] introduced formally the ontology-composition algebra within the ONION tool for ontology articulation. The authors argue against the need and possibility of constructing and maintaining a global consistent ontology. The instance-based ontology mapper GLUE, introduced by Doan and co-workers, utilizes machine learning techniques for deriving (semi)automatically assertions on the concept similarity [5]. Machine learning techniques have also been applied by Lacher and Groh in their matching tool CAIMAN [15] which is based on the instances and the documents contained in the ontological nodes. An approach combining instance-based and structural similarity measures is introduced in [20] by the author of the current paper.

The approach that we are about to present has a couple of advantages, compared to some state-of-the-art techniques. In contrast to most of the existing instance-based mapping procedures, the presented approach does not rely on instance sets intersections and can be applied for ontologies populated with entirely different document sets. The evaluation of the concepts pair-wise similarity is done at once by the help of an easy to interpret geometrical representation. This prevents us from having to evaluate m times n concept pairs for two ontologies - one with n and another with m concepts. 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 vector features are translated into a single target language. The suggested advantages significantly reduce the computational complexity of the method and increase its time efficiency.

4 From Data Analysis to Concept Mapping

The structure of two ontologies which is important for their similarity can be revealed by the help of statistical analysis methods which capture and expose information on the class separation of the ontology instances. In the current section we introduce a geometrical interpretation technique for detecting mappings among the concepts of two hierarchical ontologies by the help of a principle component analysis (PCA) and discriminant analysis (DA). The section closes with a definition of a concept similarity measure resulting from the descriptive approach.

4.1 Principle Components Analysis

PCA [11] is one of the most general data analytical methods known from descriptive statistics. It helps to extract the most essential structural information contained in a database and serves as a basis for different methods of discrimination, classification or regression. It is based on constructing new features, or principle components, by solving an Eigenvalue problem. The principle components are linear combinations of the original input variables[1] and are the new coordinates by which we represent the data. They approximate the data in the best possible way by capturing the directions of the biggest dispersion. Thus, PCA allows the representation of a multivariate data table containing thousands of variables in a lower-dimensional space (2 to 5 dimensions) by preserving and revealing the essential structural information contained in the data. In result, PCA shows what was not explicitly seen before: outliers or groups of instances are revealed; important information about the relations between variables and instances on the one hand, and in-between variable relations, on the other hand is made available.

The approach that we suggest consists in the following. We start by insisting that the ontologies have been populated following a non-hierarchical instantiation (i.e. a document is assigned to one concept only and not to all of its predecessors as well). Let D be a set of documents assigned to an ontology O with a set of concepts C. We define $l : D \mapsto C$ to be the injection which assigns to each document the label of the concept of which this document is an instance. Through l, every document is identified by its class only.

Let $O_1 := (C_1, \texttt{is\_a})$ and $O_2 := (C_2, \texttt{is\_a})$ be two hierarchical ontologies and let D_1 and D_2 be their corresponding document sets (see Section 2) of cardinalities m_1 and m_2, respectively. We can assume without loss of generality that the documents from both sets have been constructed on the same set of attributes. Let each element of the document sets has been labeled by the function l. Our goal is to find the correct mapping M_{AB} for every pair of concepts (A, B) such that $A \in C_1$ and $B \in C_2$. We produce a new dataset by taking the union of both document sets and the labels of their elements and let $D_{1,2}$ be that set. Thus we come up with a multivariate data table that contains n real variables - the dimensions of the TF/IDF vectors, and $m_1 + m_2$ observations[2] - the labeled documents from the two ontology document sets.

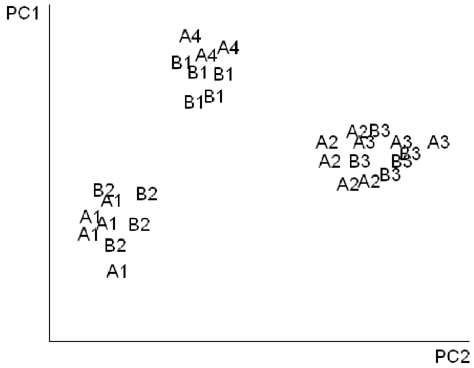


Figure 1. An example of a PCA-based concept mappings

We proceed to carry out a Principle Component Analysis on the set $D_{1,2}$. Since all our observations now live in one single space PCA will project all documents in a single principle components feature space. As we already noted, PCA shows how observations are regrouped and thus identifies the existence of classes and their relations. Naturally, all documents belonging to one single concept (no matter from which ontology) will appear to be grouped together. What is more, documents that belong to two or more different concepts from two different ontologies will also appear to be grouped together if the concepts of which they are instances are similar. What remains to do is take the labels of the documents which form one single group in the principle components projection and identify a mapping between the corresponding concepts.

An example illustrating the procedure is given on Figure 1. We see documents from concepts A1, A2, A3 and A4 from one ontology and documents from concepts B1, B2 and B3 from another. PCA shows that there are three main groups of observations. What our procedure suggests is that the documents that are grouped together in the PCA plot are instances of concepts which are to be mapped, i.e. A1 is mapped to B2, A4 - to B1, and A2 and A3 - to B3.

One straightforward problem with the proposed procedure is that principle component analysis relies on a couple of normality and linearity assumptions - too strong restrictions when dealing with ontological data. A solution to that problem is applying a non-linear version of the principle component analysis, like the one introduced in [4] or [19], based on using dot products in a feature space in terms of kernels in the input space.

[1] The term ”variable” in statistics stands for the commonly used terms ”attribute” or ”feature” in computer science. It denotes the original input variables while the term ”feature” denotes the variables that have been created out of the inputs.

[2] The term ”observations” is common for denoting the examples (or instances) in a dataset.

4.2 Discriminant Analysis

PCA finds principle components by describing as much variance of the data as possible and the case is that in practice the first components may not (and often will not) reveal the class structure that we need. Discriminant analysis, originally introduced by Fisher [7], comes in to compensate for that drawback.

Similarly to PCA, discriminant analysis is also based on constructing principle (discriminant) axes but with the explicit objective to capture the separation of the classes by minimizing their in-class variation and maximizing the distances between their means. The class information has to be included in the input data from the start. The resulting discriminant axes are again linear combinations of the input variables, where the variables with greatest weights for the construction of a given axis are the most important ones for the class separation projected on this axis. This gives rise to one of the most popular applications of DA analysis as a variable selection tool in class discrimination problems.

As in the PCA case we apply a similar geometrical approach. We take an input dataset $D_{1,2}$, as introduced above and by the help of a discriminant analysis, we identify overlaps of groups of observations. Our argumentation is the same as before: *if two (or more) classes of observations that belong to two different input ontologies appear to overlap when projected on the DA discriminant axes, the concepts of which they are instances are assigned a mapping of a similarity degree according to the size of the overlap or the distance between the classes.* Since the basic motivation of DA is to provide a proper separation of previously given classes, it is a reasonable suggestion that those classes between which DA cannot properly discriminate are similar. Applying the variable selection aspect of DA gives rise to a concept similarity measure to be introduced in the next subsection.

Finally, we note that a kernel version of Fisher's discriminant analysis handling nonlinear cases has been elaborated in [13].

4.3 A Concept Similarity Measure

Both PCA and DA project the input data on principle axes which are constructed by linear combinations of the input variables. Naturally, different linear combinations, i.e. assigning different weights to the input variables, corresponds to constructing different axes. In our approach, what we are interested in both PCA and DA analyses is the way that classes are represented and separated in a projection over one or two principle axes. For that reason, the variables which contribute at most for the construction of these axes are those that are most important for the separation of the instances projected over these axes. These variables are also said to best discriminate between the classes. In the remainder of the section we will introduce a concept similarity measure based on coinciding discriminating variables for two concepts.

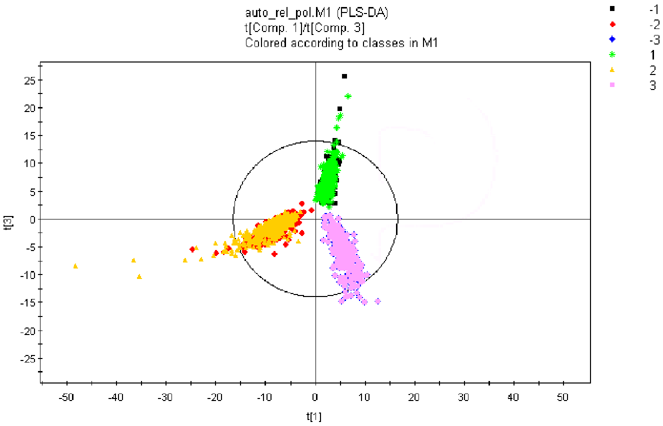


Figure 2. Discriminant Analysis plot of 6 pairwise similar classes.

Let us consider the ontology O_1 populated with documents $D_1 = \{\mathbf{d}_1^1, ..., \mathbf{d}_{m_1}^1\}$ and the ontology O_2 populated with documents $D_2 = \{\mathbf{d}_1^2, ..., \mathbf{d}_{m_2}^2\}$.

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 two classes - (1) documents that belong to the concept A and (-1) documents that do not (all the rest).

A similar data set can be acquired analogously for any concept in ontologies O_1 and O_2 and let $S^B = \{(\mathbf{d}_i^2, y_i^B)\}$, where $\mathbf{d}_i^2 \in \mathbb{R}^n$ and $i = 1, ..., m_2$ is the dataset which provides a similar separation of the instances in ontology O_2 into such that belong to the concept B and such that do not.

The measure of similarity which we are about to present is based on finding the discriminant variables in the two-class dataset S^A and comparing them to the discriminant variables found for the dataset S^B. The concepts A and B are found similar if the classes within the sets S^A and S^B are found to be separated by similar discriminant variables. Our main heuristics can be formulated as: *similar variables separate similar concepts from dissimilar ones within two ontologies*. 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 discriminant 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 first k top ordered 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 $L_k^B = \{Var_{\delta(j_1)}, Var_{\delta(j_2)}, ..., Var_{\delta(j_k)}\}$, $i, j \in (1, n)$, each of which contains the k most important variables for the separation of the instances in each corresponding ontology into such that belong to concept A, respectively B, and such that do not. The similarity of concepts A and B is defined as

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

with $sim(A, B) \in (0, 1)$. The cardinality of L_k^A and L_k^B is the same (k), for which reason the choice of a denominator of (1) is arbitrary.

5 Experiments

We carried out preliminary experiments in order to show the viability of our claims. We used data from the publicly available ”20 Newsgroups” dataset [1] which is a collection of approximately 20,000 news articles, partitioned in 20 different topics. Documents were transformed in numerical TF/IDF format by the help of RapidMiner [14] and analysis have been carried out with an already existing multivariate data analysis tool [2].

We mimicked two ontologies by taking three classes from the 20 Newsgroups dataset - Autos, Religion and Politics and splitting the instances in each class in two even parts. In that manner we constructed the pseudo ontologies O1={”Autos 1”, ”Religion 1”, ”Politics 1”} and O2={”Autos 2”, ”Religion 2”, ”Politics 2”}. We carried out a discriminant analysis and the results can be seen on Figure 2. (On the plot the positive labels correspond to the concepts of O1 and the negative ones - to the concepts in O2.) In conformity with our claims and with the semantical nature of the selected classes, the six classes appeared regrouped in three groups, where Autos 1 from O1 overlaps with Autos 2 from O2, Religion 1 overlaps with Religion 2 and Politics 1 overlaps with Politics 2.

In order to show that coinciding discriminant variables are a reliable indication of concept similarity we took the classes Autos and Religion from the 20 Newsgroups dataset. This time only the documents in Autos were split in two, mimicking two similar concepts and ”Religion” was kept as a class which plays the role of a complement of the concepts ”Autos 1” in ontology O1 and ”Autos 2” in the ontology O2. We carried out a discriminant analysis on the set of labeled documents from the three introduced classes. As we can see on Figure 3, the two autos classes (labeled by 1 and 2) appear close to each other, almost completely overlapping, while ”Religion” (labeled by -1) remains clearly separated. The interpretation is that ”Autos 1” and ”Autos 2” are similar to each other and dissimilar from ”Religion”. In order to reinforce this finding and justify the similarity measure introduced in the previous section, we carried out two additional discriminant analysis, this time focusing on extracting the discriminant variables:

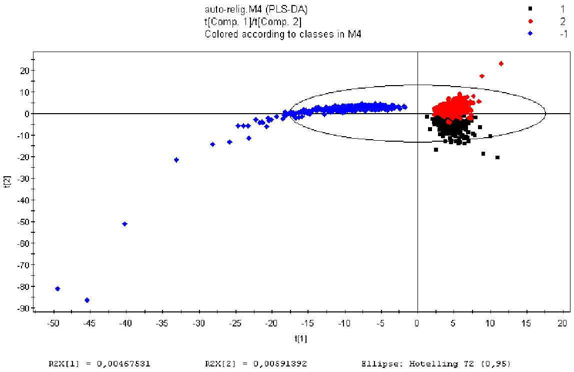
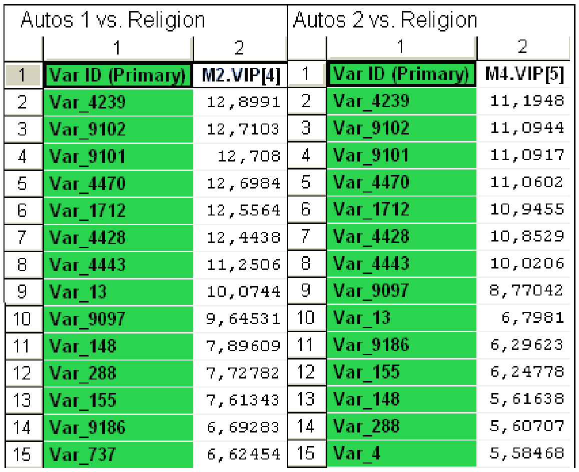


Figure 3. A DA scatter plot of the classes Autos 1, Autos 2 and Religion.



Autos 1 vs. Religion			Autos 2 vs. Religion		
	1	2		1	2
1	Var ID (Primary)	M2.VIP[4]	1	Var ID (Primary)	M4.VIP[5]
2	Var_4239	12,8991	2	Var_4239	11,1948
3	Var_9102	12,7103	3	Var_9102	11,0944
4	Var_9101	12,708	4	Var_9101	11,0917
5	Var_4470	12,6984	5	Var_4470	11,0602
6	Var_1712	12,5564	6	Var_1712	10,9455
7	Var_4428	12,4438	7	Var_4428	10,8529
8	Var_4443	11,2506	8	Var_4443	10,0206
9	Var_13	10,0744	9	Var_9097	8,77042
10	Var_9097	9,64531	10	Var_13	6,7981
11	Var_148	7,89609	11	Var_9186	6,29623
12	Var_288	7,72782	12	Var_155	6,24778
13	Var_155	7,61343	13	Var_148	5,61638
14	Var_9186	6,69283	14	Var_288	5,60707
15	Var_737	6,62454	15	Var_4	5,58468

Figure 4. Variables discriminating between {Autos 1 and Religion} and {Autos 2 and Religion}

(1) Find the important variables for the separation of ”Autos 1” and ”Religion”;

(2) Find the important variables for the separation of ”Autos 2” and ”Religion”.

The results, presented in the table on Figure 4 showed that the list of the variables discriminating between ”Autos 1” and ”Religion” is very similar (almost identical) to the list of the variables discriminating between ”Autos 2” and ”Religion”. (On the figure, VIP stands for a score coefficient calculated on the basis of the contribution of a single variable to the construction of the discriminant axes.) Applying the similarity measure (1) leads to identifying a similarity mapping between the concepts ”Autos 1” and ”Autos 2”.

6 Conclusion

The paper describes an approach for selecting sets of potential concept mappings from two different ontologies. The method is based on geometrical interpretation of the structural information contained in the instance sets of both ontologies. In addition to the proposed geometrical approach we define a novel concept similarity measure based on discriminant variables. Apart from a self-standing concept similarity detection tool, the described approach can be used as a support for an independent concept similarity measure, indicating groups of potentially similar concepts. The described procedure can be applied to mapping heterogeneous web directories.

References

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

[2] http://www.umetrics.com/

[3] T. Berners-Lee, J. A. Hendler, O. Lassila. The Semantic Web, In: *Scientific American,* 284(5):34-43, 2001.

[4] G. Blanchard, P. Massart, R. Vert, L. Zwald. Kernel Projection Machine: a New Tool for Pattern Recognition, In *Proceedings of NIPS,* 2004.

[5] 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.

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

[7] R.A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7: 179-188, 1936.

[8] T. R. Gruber. Towards Principles for the Design of Ontologies Used for Knowledge Sharing, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, 1993.

[9] 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.

[10] 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.

[11] I.T. Jolliffe. *Principle Component Analysis*. Springer-Verlag, New York, New York 1986.

[12] Y. Kalfoglou, M. Schorlemmer. Ontology mapping: the state of the art, *Knowl. Eng. Rev.*, 18(1):1–31, 2003.

[13] 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.

[14] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, T. Euler. YALE: Rapid Prototyping for Complex Data Mining Tasks. *KDD ’06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 935–940, 2006.

[15] 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.

[16] 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.

[17] N. Noy, M. Musen. The PROMPT suite: Interactive tools for ontology merging and mapping, *Noy, N., Musen, M.: The PROMPT suite: Interactive tools for ontology merging and mapping. Technical report, SMI, Stanford University, CA, USA (2002)*, 2002.

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

[19] B. Scholkopf, A.J. Smola, K.-R. Muller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Computation* vol. 10, 1299-1319, 1998.

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