Fuzzy Clustering for Categorical Spaces
An Application to Semantic Knowledge Bases

Nicola Fanizzi, Claudia d'Amato, and Floriana Esposito

Dipartimento di Informatica, Università degli studi di Bari
Campus Universitario, Via Orabona 4, 70125 Bari, Italy
{fanizzi,claudia.damato,esposito}@di.uniba.it

Abstract. A multi-relational clustering method is presented which can be applied to complex knowledge bases storing resources expressed in the standard Semantic Web languages. It adopts effective and language-independent dissimilarity measures that are based on a finite number of dimensions corresponding to a committee of discriminating features(represented by concept descriptions). The clustering algorithm expresses the possible clusterings in tuples of central elements (medoids, w.r.t. the given metric) of variable length. It iteratively adjusts these centers following the rationale of fuzzy clustering approach, i.e. one where the membership to each cluster is not deterministic but rather ranges in the unit interval. An experimentation with some ontologies proves the feasibility of our method and its effectiveness in terms of clustering validity indices.

1 Clustering in Complex Domains

Recently, multi-relational learning methods are being devised for knowledge bases in the Semantic Web (henceforth SW), expressed in the standard representations. Indeed, the most burdensome related maintenance tasks, such as *ontology construction*, *refinement* and *evolution*, demand such automatization also to enable further SW applications.

In this work, we investigate on unsupervised learning for knowledge bases expressed in such standard languages. In particular, we focus on the problem of clustering semantically annotated resources. The benefits of clustering can be manifold. Clustering annotated resources enables the definition of new emerging concepts (*concept formation*) on the grounds of the concepts defined in a knowledge base; supervised methods can exploit these clusters to induce new concept definitions or to refining existing ones (*ontology evolution*); intensionally defined groupings may speed-up the task of approximate search and *discovery* [1]; a clustering may also suggest criteria for *ranking* the retrieved resources based on the distance from the cluster centers.

Most clustering methods are based on the application of similarity (or density) measures defined over a set of attributes of the domain objects. Classes of objects are taken as collections that exhibit low interclass similarity (density) and high intraclass similarity (density). Few methods are able to take into account some form of *background knowledge* that could characterize object configurations by means of global concepts and semantic relationships [2].

Specific approaches designed for terminological representations (*Description Logics* [3], henceforth DLs), have been introduced [4, 5]. These logic-based clustering methods

J. Rauch et al. (Eds.): ISMIS 2009, LNAI 5722, pp. 161–170, 2009.

were devised for some specific DL languages of limited expressiveness. The main drawback of these methods is that they are language-dependent, which prevents them to scale to the standard SW representations that are mapped on complex DLs. Moreover, purely logic methods can hardly handle noisy data while distance-based ones may be more robust. Hence, from a technical viewpoint, suitable measures for concept representations and their semantics are to be devised. A further theoretical problem comes from the increased indeterminacy determined by the *Open-World semantics* that is adopted on the knowledge bases, differently from the *Closed-World* semantics which is more generally adopted in other contexts (e.g. databases).

These problems motivate the investigation on similarity-based clustering methods which can be more noise-tolerant and language-independent. Specifically, the extension of distance-based techniques is proposed, which can cope with the standard SW representations and profit by the benefits of a randomized search for optimal clusterings. Indeed, the method is intended for grouping similar resources w.r.t. a notion of similarity, coded in a distance measure, which fully complies with the semantics of knowledge bases expressed in DLs. The individuals are gathered around cluster centers according to their distance. The choice of the best centers (and their number) is performed through a fuzzy membership approach [6].

Although some structural dissimilarity measures have been proposed for some specific DLs of fair expressiveness [1], they are still partly based on structural criteria which makes them fail to fully grasp the underlying semantics and hardly scale to more complex ontology languages such as those backing the OWL ontology language[1]. Therefore, we have devised a family of semi-distance measures for semantically annotated resources, which can overcome the aforementioned limitations [7, 8]. Following the criterion of semantic discernibility of individuals, a family of measures is derived that is suitable for a wide range of languages since it is merely based on the discernibility of the input individuals with respect to a fixed committee of features represented by a set of concept definitions. Hence, the new measures are not absolute, they rather depend on the knowledge base they are applied to. Thus, also the choice of good feature may deserve a preliminary optimization phase, which can be performed by means of a randomized search procedures [8].

In the target setting, the notion of *centroid* characterizing distance-based algorithms for numeric representations descending from K-MEANS [9], is replaced by the notion of *medoids* as cluster prototypes which fit better categorical representations [10]. Differently from these deterministic approaches, the proposed clustering algorithm employs a notion of fuzzy membership w.r.t. the current medoids computed according to the measure mentioned above. On each iteration, the choice of medoids evolves by adjusting the membership probability w.r.t. each medoid.

The paper is organized as follows. Sect. 2 presents the basics of the target representation and the semantic similarity measures adopted. This algorithm is presented and discussed in Sect. 3. We report in Sect. 4 an experiment aimed at assessing the validity of the method on some ontologies available in the Web. Conclusions and extensions are finally examined in Sect. 6.

[1] `http://www.w3.org/TR/owl-guide/`

2 Metrics for DL Representations

2.1 Preliminaries on the Representation

In the following, we assume that resources, concepts and their relationship may be defined in terms of a generic ontology language that may be mapped to some DL language with the standard model-theoretic semantics (see the DLs handbook [3] for a thorough reference). As mentioned in the previous section, one of the advantages of our method is that it does not depend on a specific language for semantic annotations based on DLs. However the implementation applies to OWL-DL knowledge bases.

In the reference DL framework, a *knowledge base* $\mathcal{K} = \langle \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle$ contains a *TBox* $\mathcal{T}$, an *RBox* $\mathcal{R}$ and an *ABox* $\mathcal{A}$. $\mathcal{T}$ is a set of concept definitions: $C \equiv D$, where C is the atom denoting the defined concept and D is a DL concept description specified by the application of the language constructors to *primitive concepts* and *roles*. The *RBox* $\mathcal{R}$ contains similar axioms for specifying new roles by means of proper constructors. The complexity of such definitions depends on the specific DL language. $\mathcal{A}$ contains *assertions* (ground facts) on *individuals* (domain objects) concerning the current world state, namely $C(a)$ (*class-membership*), a is an instance of concept C, and $R(a, b)$ (*relations*), a is R-related to b. The set of the individuals referenced in the assertions ABox $\mathcal{A}$ is usually denoted with $\mathsf{Ind}(\mathcal{A})$. Each individual can be assumed to be identified by a constant (or its own URI in OWL-DL), however this is not bound to be a one-to-one mapping (*unique names assumption*).

A set-theoretic semantics is generally adopted with these representations, with interpretations $\mathcal{I}$ which map each concept description C to a subset of a domain of objects (*extension*) $C^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ and each role description R to $R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$. This allows the formation of complex hierarchies of concept/roles.

In this context the most common inference is the computation of the *subsumption* relationship between concepts: given two concept descriptions C and D, D *subsumes* C, denoted by $C \sqsubseteq D$, iff for every interpretation $\mathcal{I}$ it holds that $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$. The interpretations of interest are those that satisfy all axioms in the knowledge base $\mathcal{K}$, i.e. its *models*. Model-theoretic entailment will be denoted with $\models$.

Several other inference services are provided by the standard automated reasoners. Like all other instance-based methods, the measures proposed in this section require performing *instance-checking*, which amounts to determining whether an individual, say a, belongs to a concept extension, i.e. whether $\mathcal{K} \models C(a)$ holds for a certain concept C. In the simplest cases (primitive concepts) instance-checking requires simple ABox lookups, yet for defined concepts the reasoner may need to perform a number of inferences. It is worthwhile to recall that the *Open World Assumption* (OWA) is made. Thus, differently from the standard database framework, reasoning procedures might be unable to ascertain the class-membership or non-membership. Hence one has to cope with this form of uncertainty.

2.2 Comparing Individuals within Ontologies

In distance-based cluster analysis, a function for measuring the (dis)similarity of individuals is needed. It can be observed that individuals do not have a syntactic structure

that can be compared. This has led to lifting them to the concept level before comparing them [1] (resorting to the approximation of the *most specific concept* of an individual w.r.t. the ABox [3]).

Inspired from some techniques for distance construction and *Multi-dimensional Scaling* [6, 11], we have proposed the definition of totally semantic distance measures for individuals in the context of a knowledge base which is also able to cope with the OWA. On a semantic level, similar individuals should behave similarly with respect to the same concepts. We have introduced a novel measure, which is based on the idea of comparing their semantics along a number of dimensions represented by a committee of concept descriptions. Thus, the rationale of the new measure is to compare individuals on the grounds of their behavior w.r.t. a given collection of concept descriptions, say $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$, which stands as a group of discriminating *features* expressed in the considered DL language.

The general form of the family of dissimilarity measures for individuals inspired to the Minkowski's distances (L_p) can be defined as follows [7, 8]:

Definition 2.1 (family of dissimilarity measures). *Let $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ be a knowledge base. Given a set of concept descriptions $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$ and a normalized vector of related weights $\boldsymbol{w}$, a family of* dissimilarity measures *$\{d_p^{\mathsf{F}}\}_{p \in \mathbb{N}}$, with $d_p^{\mathsf{F}} : \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A}) \mapsto [0, 1]$, is defined as follows:*

$$\forall a, b \in \mathsf{Ind}(\mathcal{A}) \qquad d_p^{\mathsf{F}}(a, b) := \left[\sum_{i=1}^{m} \mid w_i \cdot (\pi_i(a) - \pi_i(b)) \mid^p \right]^{\frac{1}{p}}$$

where the projection function vector *$\boldsymbol{\pi}$ is defined $\forall i \in \{1, \ldots, m\}$*

$$\forall a \in \mathsf{Ind}(\mathcal{A}) \qquad \pi_i(a) = \begin{cases} 1 & \mathcal{K} \models F_i(a) \qquad (\textit{alt. } F_i(x) \in \mathcal{A}) \\ 0 & \mathcal{K} \models \neg F_i(a) \qquad (\textit{alt. } \neg F_i(x) \in \mathcal{A}) \\ 1/2 & \textit{otherwise} \end{cases}$$

Note that the measure is efficiently computed when the feature concepts F_i are such that the KBMS can directly infer the truth of the assertions $F_i(a)$, ideally $\forall a \in \mathsf{Ind}(\mathcal{A})$: $F_i(a) \in \mathcal{A}$. This is very important for the measure integration in algorithms which massively use them, such as all instance-based methods. The presented method can be regarded as a form of propositionalization [12].

The given definition of the projection functions is basic. The case of $\pi_i(a) = 1/2$ corresponds to the case when a reasoner cannot give the truth value for a certain membership query. This is due to the open-world semantics adopted in this context.

An intermediate value is just a raw (uniform) estimate of the uncertainty related to the single feature. By properly assigning the weights to vector $\boldsymbol{w}$ it is possible to obtain a better measure which reflects the available knowledge [11].

3 Fuzzy Clustering in Complex Domains

The schemata of many similarity-based clustering algorithms (see [9] for a survey) can be adapted to more complex settings like the one of interest for this work, especially when similarity or similarity measures are available.

We focus on a generalization of distance-based methods adopting the notion of prototypes as cluster centers [10]. The method implements a fuzzy clustering scheme [6], where the representation for the clusterings that are iteratively adjusted is made up of tuples of prototypical individuals for the various clusters but, differently from the k-MEANS the membership of the instances to the various clusters is probabilistic rather than deterministic.

The algorithm searches the space of possible clusterings of individuals, optimizing a fitness function L based on the relative discernibility of the individuals of the different clusters (inter-cluster separation) and on the intra-cluster similarity measured in terms of the d_p^{F} pseudo-metric. Considered a set of cluster centers (prototypes) $\{\mu_1, \ldots, \mu_k\}$, a notion of graded membership of an individual x_i w.r.t. a given cluster C_j is introduced ranging in $[0, 1]$. This corresponds to computing the probability $P(C_j|x_i, \theta)$.

The objective function to be minimized can be written:

$$L = \sum_{i=1}^{N} \sum_{j=1}^{k} (P(C_j|x_i, \theta))^b d(x_i, \mu_j)$$

Its minima are found solving the equations involving the partial derivatives w.r.t. the medoids $\partial L/\partial \mu_j = 0$ and of the probability $\partial L/\partial \hat{P}_j = 0$, yielding:

$$\mu_j = \frac{\sum_i (P(C_j|x_i))^b \cdot x_i}{\sum_i (P(C_j|x_i))^b} \qquad \forall j \in \{1, \ldots, k\} \tag{1}$$

and

$$P(C_j|x_i) = \frac{(1/d_{ij})^{\frac{1}{b-1}}}{\sum_r (1/d_{ir})^{\frac{1}{b-1}}} \qquad \forall i \in \{1, \ldots, N\} \quad \forall j \in \{1, \ldots, k\} \tag{2}$$

where $d_{ij} = d(x_i, \mu_j)$.

In a categorical setting, the notion of *medoid* was introduced [10, 9] for categorical feature-spaces w.r.t. some distance measure. Namely, the medoid of a group of individuals is the individual that has the minimal average distance w.r.t. the others. Formally:

Definition 3.1 (medoid). *Given a set of individuals S and a dissimilarity measure d, the medoid of the set is defined:*

$$\mu_S = \text{medoid}(S) := \operatorname*{argmin}_{a \in S} \frac{1}{|S|} \sum_{b \in S} d(a, b) \tag{3}$$

In the setting of interest, the prototypes are not numerical tuples but actual individuals (medoids). Eqs. 1 and 3 may be summed up in a single one as follows:

$$\mu_j = \operatorname*{argmin}_{a \in C_j} \sum_{b \in C_j} d(a, b) \cdot P(C_j|a) \qquad \forall j \in \{1, \ldots, k\} \tag{4}$$

i.e. the medoids are determined by the individuals minimizing the distance to the other members of the cluster, weighted by their membership probability. Finally, a specific similarity measure for individuals like those defined in the previous section is needed: $d = d_p^{\mathsf{F}}$ (for some F and p).

clustering FUZZY-k-MEDOIDS(k, individuals, maxIterations)
input: k: required number of clusters;
individuals: individuals to be clustered;
maxIterations: maximum number of iterations;
output: clustering: set of clusters

begin
Initialize iteration $\leftarrow 0$, random prototypes $M = \{\mu_j\}_{j=1}^{k}$
Initialize uniform probabilities $P(C_j|x_i)$, for $i = 1, \ldots, N$, $j = 1, \ldots, k$
repeat
For each $a \in$ individuals:
$t \leftarrow \text{argmin}_{j=1,\ldots,k}\, d(a, \mu_j)$
$C_t \leftarrow C_t \cup \{a\}$
re-compute prototypes $M = \{\mu_j\}_{j=1}^{k}$ according to eq. (4)
re-compute all probabilities $P(C_j|x_i)$, using eq. (1)
normalize the probabilities, for $i = 1, \ldots, N$
++iteration
until convergence **or** iteration = maxIterations
return $\{C_j\}_{j=1,\ldots,k}$
end

Fig. 1. The fuzzy clustering algorithm for categorical metric spaces

Fig. 1 reports a sketch of the FUZZY k-MEDOIDS algorithm. Note that the algorithm requires the number of clusters k as a parameter.

The representation of centers through medoids has two advantages. First, it presents no limitations on attributes types, and, second, the choice of medoids is dictated by the location of a predominant fraction of points inside a cluster and, therefore, it is less sensitive to the presence of outliers. This robustness is particularly important in the common context that many elements do not belong exactly to any cluster, which may be the case of the membership in DL knowledge bases, which may be not ascertained given the OWA. Algorithms where prototypes are represented by centroids, which are weighted averages of points within a cluster work conveniently only with numerical attributes and can be negatively affected even by a single outlier. An algorithm based on medoids allows for a more flexible definition of similarity. Many clustering algorithms work only after transforming symbolic into numeric attributes.

4 Evaluation

The clustering algorithm has been evaluated with an experimentation on various knowledge bases selected from standard repositories. The option of randomly generating assertions for artificial individuals was discarded for it might have biased the procedure. Only populated ontologies (which may be more difficult to find) were suitable for the experimentation.

A number of different knowledge bases represented in OWL were selected from various sources (the Protégé library[2] and the Swoogle[3] search engine were used), namely: FINITESTATEMACHINES (FSM), SURFACEWATERMODEL (SWM), TRANSPORTATION, WINE, NEWTESTAMENTNAMES (NTN), FINANCIAL, the BioPax glycolysis ontology (BIOPAX), and one of the ontologies generated by the Lehigh University Benchmark (LUBM). Table 1 summarizes salient figures concerning these ontologies.

[2] `http://protege.stanford.edu/plugins/owl/owl-library`
[3] `http://swoogle.umbc.edu`

Table 1. Ontologies employed in the experiments

Ontology	DL language	#concepts	#object props.	#data props.	#individuals
FSM	$\mathcal{SOF}(D)$	20	10	7	37
SWM	$\mathcal{ALCOF}(D)$	19	9	1	115
LUBM	$\mathcal{ALR}_{+}\mathcal{HI}(D)$	43	7	25	118
WINES	$\mathcal{ALCIO}(D)$	112	9	10	149
BIOPAX	$\mathcal{ALCIF}(D)$	74	70	40	323
NTN	$\mathcal{SHIF}(D)$	47	27	8	676
FINANCIAL	$\mathcal{ALCIF}$	60	16	0	1000

In the computation of the proximity matrix (the most time-consuming operation) all named concepts in the knowledge base have been used for the committee of features, thus guaranteeing meaningful measures with high redundancy. The squared version of the measure has been adopted (d_2^F) with uniform weights. The PELLET reasoner[4] (ver. 2.0rc4) was employed to perform the inferences that were necessary to compute the proximity matrices. The experimentation consisted of 50 runs of the algorithm per knowledge base. Each run took from a 1 to 5 mins on a QuadCore (2Gb RAM) Linux box, depending on the specific ontology processed. The indices which were chosen for the experimental evaluation of the outcomes were the following: an alternative R-Squared index (ranging in $[0, 1]$) [13] adopting medoids as cluster centers, Hubert's normalized Γ index [13] and the average Silhouette index [10], both ranging in $[-1, 1]$, with 1 indicating the best performance. We also considered the average number of clusters resulting from the runs on each knowledge base. It is also interesting to compare this number to the one of the primitive and defined concepts in each ontology (see Table 1, rightmost column).

For a comparison w.r.t. a different (stochastic) clustering procedure which is applicable to the same datasets employed in previous works (tables can be found in [8]), we opted for the average number of clusters found during the runs of the algorithm. Table 2 reports the average outcomes of these experiments. The table shows that the algorithm is quite stable in terms of all indices, as testified by the low variance of the results, despite its inherent randomized nature. As such, the optimization procedure does not seem to suffer from being caught in local minima.

Hubert's normalized Γ index measures both compactness and separation of the resulting clusters w.r.t. the proximity matrix. Results are generally good for the various ontologies. The R-Squared average values denote a good degree of separation between the various clusters. We may interpret the outcomes observing that clusters present a high degree of compactness. It should also pointed out that flat clustering penalizes separation as the concepts in the knowledge base are seldom declared to be disjoint. Rather, they naturally tend to form subsumption hierarchies. As for the average Silhouette index the performance of the algorithm is generally very good with a slight degradation with the increase of individuals taken into account. Besides, note that the largest knowledge base (in terms of its population) is also the one with the maximal number of concepts which provided the features for the metric. Surprisingly, the number of clusters is limited w.r.t. the number of concepts in the KB, suggesting that many individuals gather around a restricted subset of the concepts, while the others are only

[4] `http://clarkparsia.com/pellet`

Table 2. Results of the experiments with the FUZZY k-MEDOIDS algorithm

Ontology	Hubert's Γ	R-Squared	Silhouette	#clusters
FSM	.51 (±8.29e-2) [.39,.72]	.81 (±4.98e-2) [.74,.92]	.81 (±3.64e-2) [.74,.89]	13
SWM	.77 (±4.99e-2) [.63,.73]	.85 (±1.74e-2) [.81,.89]	.88 (±6.49e-2) [.81,.95]	14
LUBM	.60 (±9.14e-2) [.48,.75]	.51 (±1.09e-1) [.31,.69]	.85 (±2.05e-2) [.75,.90]	12
WINE	.32 (±4.30e-2) [.26,.41]	.98 (±6.56e-4) [.982,.985]	.88 (±1.42e-2) [.84,.90]	78
BIOPAX	.59 (±7.77e-4) [.45,.77]	.62 (±7.00e-2) [.45,.78]	.88 (±1.46e-2) [.85,.92]	16
NTN	.86 (±2.00e-2) [.76,.88]	.83 (±3.35e-2) [.65,.88]	.93 (±1.77e-2) [.90,.95]	35
FINANCIAL	.44 (±1.36e-2) [.42,.46]	.46 (±2.26e-2) [.43,.45]	.89 (±3.26e-2) [.85,.92]	27

complementary (they can be used to discern the various individuals). Such subgroups may be detected extending our method to perform hierarchical clustering.

5 Hierarchical Clustering

Some natural extensions may be foreseen for the presented algorithm. One regards upgrading the algorithm so that it may build *hierarchical* clusterings levelwise in order to produce (or reproduce) terminologies possibly introducing new concepts elicited from the ontology population. Hierarchical clustering methods may adopt agglomerative (*clumping*) or divisive (*splitting*) approaches and usually require distance functions for calculating distance between clusters.

Given the algorithm presented in section 1, it appears natural to focus on divisive methods. Whereas agglomerative clustering begins with each element a cluster and then combines clusters using a distance measure, divisive hierarchical clustering begins with one cluster and then continually breaks these clusters into smaller and smaller clusters until a stopping criterion is met (no quality improvement or singleton clusters reached). The clusters at each level are examined and the one containing objects that are the farthermost according to the metric are broken apart.

The hierarchical extension of the algorithm implements a divisive method, starting with one universal cluster grouping all instances. Iteratively, it creates new clusters by applying the FUZZY k-MEDOIDS to the worst cluster and this continues until a stopping criterion is met, so that finally a *dendrogram* is produced. Fig. 2 reports a sketch of the algorithm. It essentially consists of a loop that computes a new level of the dendrogram until the stopping criterion is met; the inner call to FUZZY k-MEDOIDS returns a clustering of one cluster at the current level.

On each level, the worst cluster is selected (call to the SELECTWORSTCLUSTER function) on the grounds of its quality, e.g. the one endowed with the least average inner similarity (or cohesiveness). This cluster is candidate to being split. The partition is constructed by calling FUZZY k-MEDOIDS on the worst cluster (worstCluster). In the end, the candidate cluster is replaced by the newly found parts at the next level of the dendrogram (call to the REPLACE function).

```
clusterVector HIERARCHICALFCM(allIndividuals, params)
input   allIndividuals: list of individuals
        params: other parameters for FUZZY-k-MEDOIDS
output  clusterVector: array of lists of clusters

begin
level ← 0;
clusterVector[1] ← allIndividuals;
repeat
    level ← level + 1;
    worstCluster ← SELECTWORSTCLUSTER(clusterVector[level]);
    newClusters ← FUZZY-k-MEDOIDS(worstCluster,params);
    clusterVector[level+1] ← REPLACE(worstCluster,newClusters,clusterVector[level]);
until STOPCRITERION(clusterVector[level+1]);
return clusterVector
end
```

Fig. 2. The HIERARCHICAL FUZZY k-MEDOIDS algorithm

Two criteria have not been entirely specified: the function that determines the cluster quality and the stopping condition. As regards the cluster quality, there is a plethora of choices in the literature [14, 9]. Some of these functions have been recalled in section 4, for they determine validity measures. For example, the extent of a cluster diameter can be considered as a criterion for deciding which cluster has to be split. Alternatively, one may consider all clusters as candidates to the split, perform them and then evaluate the resulting new clustering level using the validity measures referred above. Although more computationally costly this may allow considering inter-cluster separation in the splitting criterion. As concerns the stopping criterion one may simply consider a maximum number of clusters to be produced. Again, a more costly way to determine the criterion would involve an evaluation of the gain yielded by a further level $l + 1$, in terms of the validity measure of choice (vm): $\text{gain}(l + 1) = \text{vm}(\mathsf{clustering}[l + 1]) - \text{vm}(\mathsf{clustering}[l])$. An insufficient or even negative improvement of the clustering quality (e.g. w.r.t. some given threshold) may determine the halting condition for the algorithm.

Alternative divisive methods based on hierarchical extensions of the PARTITION-AROUND-MEDOIDS algorithm have been considered [7].

6 Concluding Remarks

This work has presented a framework for fuzzy clustering that can be applied to standard multi-relational representations adopted for knowledge bases in the SW context. Its intended usage is for discovering interesting groupings of semantically annotated resources and can be applied to a wide range of concept languages. Besides, the induction of new concepts may follow from such clusters [7], which allows for accounting for them from an intensional viewpoint. In this paper we have also presented a possible extension to producing hierarchical clustering. A further natural extension of the clustering algorithm is towards incrementality.

The method exploits a dissimilarity measure that is based on the underlying resource semantics w.r.t. a number of dimensions corresponding to a committee of features represented by a group of concept descriptions in the language of choice. The algorithm is an extension of distance-based clustering procedures employing medoids as cluster

prototypes so to deal with complex representations of the target context. The distance measure may also serve as a ranking criterion.

As regards the optimization of the pseudo-metric, a promising research line, for extensions to matchmaking, retrieval and classification, is *retrieval by analogy* [1]: a search query may be issued by means of prototypical resources; answers may be retrieved based on local models (intensional concept descriptions) for the prototype constructed (on the fly) based on the most similar resources (w.r.t. some similarity measure). The presented algorithm may be the basis for the model construction activity.

References

[1] d'Amato, C., Fanizzi, N., Esposito, F.: Analogical reasoning in description logics. In: da Costa, P.C.G., d'Amato, C., Fanizzi, N., Laskey, K.B., Laskey, K.J., Lukasiewicz, T., Nickles, M., Pool, M. (eds.) URSW 2005 - 2007. LNCS (LNAI), vol. 5327, pp. 330–347. Springer, Heidelberg (2008)

[2] Kirsten, M., Wrobel, S.: Relational distance-based clustering. In: Page, D.L. (ed.) ILP 1998. LNCS, vol. 1446, pp. 261–270. Springer, Heidelberg (1998)

[3] Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P. (eds.): The Description Logic Handbook. Cambridge University Press, Cambridge (2003)

[4] Kietz, J.-U., Morik, K.: A polynomial approach to the constructive induction of structural knowledge. Machine Learning 14(2), 193–218 (1994)

[5] Fanizzi, N., Iannone, L., Palmisano, I., Semeraro, G.: Concept formation in expressive description logics. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 99–110. Springer, Heidelberg (2004)

[6] Duda, R., Hart, P., Stork, D.: Pattern Classification, 2nd edn. Wiley, Chichester (2001)

[7] Fanizzi, N., d'Amato, C., Esposito, F.: Conceptual clustering for concept drift and novelty detection. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 318–332. Springer, Heidelberg (2008)

[8] Fanizzi, N., d'Amato, C., Esposito, F.: Evolutionary conceptual clustering based on induced pseudo-metrics. Semantic Web Information Systems 4(3), 44–67 (2008)

[9] Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Computing Surveys 31(3), 264–323 (1999)

[10] Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, Chichester (1990)

[11] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. In: Data Mining, Inference, and Prediction. Springer, Heidelberg (2001)

[12] Kramer, S., Lavrač, N., Flach, P.: Propositionalization approaches to relational data mining. In: Džeroski, S., Lavrač, N. (eds.) Relational Data Mining, pp. 262–286. Springer, Heidelberg (2001)

[13] Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Information Systems 17(2-3), 107–145 (2001)

[14] Bezdek, J.C., Pal, N.R.: Some new indexes of cluster validity. IEEE Transactions on Systems, Man, and Cybernetics 28(3), 301–315 (1998)