International Journal of Computational Intelligence Systems, Vol. 9, No. 5 (2016) 876-887

Ontology Knowledge Mining for Ontology Alignment

Rihab Idoudi[1,2] , Karim Saheb Ettabaa[2], Basel Solaiman[2], Kamel Hamrouni[1]

[1] *Université Tunis ElManar, Ecole Nationale d'Ingénieurs de Tunis, Tunis, 1200, Tunisie*
[2] *Telecom Bretagne, ITI Laboratory 29238, Brest, France*

Received 19 November 2015

Accepted 8 May 2016

Abstract

As the ontology alignment facilitates the knowledge exchange among the heterogeneous data sources, several methods have been introduced in literature. Nevertheless, few of them have been interested in decreasing the problem complexity and reducing the research space of correspondences between the input ontologies.This paper presents a new approach for ontology alignment based on the ontology knowledge mining. The latter consists on producing for each ontology a hierarchical structure of fuzzy conceptual clusters, where a concept can belong to several clusters simultaneously. Each level of the hierarchy reflects the knowledge granularity degree of the knowledge base in order to improve the effectiveness and speediness of the information retrieval. Actually, such method allows the knowledge granularity analyze between the ontologies and facilitates several ontology engineering techniques. The ontology alignment process is performed iteratively over the produced hierarchical structure of the fuzzy clusters using semantic techniques. Once the correspondent clusters are identified, we consider both syntactic and structural characteristics of their correspondent entities. The proposed approach has been tested over the OAEI benchmark dataset and some real mammographic ontologies since this work is a part of CMCU project for Mammographic images analysis for Assistance Diagnostic Breast Cancer. The system performs good results in the terms of precision and recall with respect to other alignment system.

Keywords: knowledge mining, Hierarchical Fuzzy clustering, Ontology Alignment, Similarity techniques.

1. Introduction

Ontology, as it represents a mean to formalize the domain knowledge, has become the enabler of the knowledge exchange between the heterogeneous data sources. However, for a specific domain, several ontologies have been developed independently by different communities and with distinct perspectives and/or objectives [1]. In practice, it is crucial to employ well specific parts from the offered ontologies to accomplish the best results of knowledge sharing. Hence, in order to achieve the semantic interoperability among the domain ontologies, it is required to discover correspondences across these knowledge bases. This ontology engineering technique is called the ontology alignment. However, for large and voluminous ontologies, where the research space is noticeably huge, the correspondences establishment becomes more complex and the effectiveness of most of these systems underperforms in terms of execution time, allocated memory size or mappings results precision [2]. In fact, during an alignment process, most of the existing approaches compare the individual couples of entities using one or more similarity technique and then the

results of these techniques are aggregated using a variety of aggregation strategies, such as, the system proposed in [3]. The latter uses three different matching strategies (name-based, metadata-based and instance-based) whose results are then filtered and combined. However, the issue arises when ontologies are voluminous with hundreds or thousands of concepts and the alignment task turns out to be hard to handle with multi-attributes entities. In the perspective of performing both speediness and effectiveness of the ontology alignment process, some researchers have tackled the problem of scalability with the use of the clustering algorithms. Such technique aims to reduce the research space of correspondences between the ontologies' entities to be aligned. In [4], a scalable solution, called Falcon-AO, for matching large ontologies has been proposed, the process starts with partitioning the ontology's entities to construct a set of small clusters. The partitioning method is based on the structural proximities between the concepts and properties using one type of relationships which is the subsumption relations ('is-a'). Then, it constructs disjoint blocks out of these clusters. In the next step, the alignment process, parses both sets of produced clusters of the ontologies and exploits the whole cluster's information to determine the similar clusters pairs having the higher proximity. This proximity is based on anchors (shared entities). The more these clusters share anchors, the more similar they are. The COMA++ system presented in [5] consists on partitioning large ontologies by using relatively simple heuristic rules. It starts by transforming ontologies into graphs. Then, clustering algorithm is applied to partition the graphs into disjoint clusters. Contrarily to the Falcon –AO system, the aligning process is based only on the partitions' roots to determine similar clusters. The use of limited information about the cluster may result in less alignment quality. To dependently cluster ontologies, TaxoMap [6] uses a co-clustering technique, the system determines similar modules through the clustering process. The system provides one-to-many mappings between single concepts. In [7], the author proposed a clustering approach based on structural nodes similarity. Therefore, each cluster of the source ontology has to be aligned with only one subset of the target ontology. In [8], the approach starts by anchoring, a pair of "look-alike"neighbors concepts to be aligned. The method outputs a set of alignments between concepts within semantically similar subsets. The authors in [9] address the problem of aligning large class hierarchies by introducing a partition-based block approach. The process is based on predefined anchors and uses structural and linguistic similarities to partition class hierarchies into small blocks. A structural clustering method based on network analysis was proposed in [10]. The latter produces, in a consuming time, an important number of too small modules (which may affect the concept's overall context). Although, those approaches contribute slightly in enhancing the alignment results they suffer from several limitations. Therefore, a generic approach called FHCbM (Fuzzy Hierarchical clustering based method) based on the ontology knowledge mining is proposed to address the challenge of the increased concepts sets size to be treated. Contrarily to the existing methods which produce flat and crisp clusters, this paper makes the following contributions:

- The ontological structure is reorganized through a hierarchical structure of fuzzy conceptual clusters. Such reorganization allows the knowledge granularity levels analyze as well as the ontology alignment process enhancing by reducing the problem complexity.

- The use of fuzzy clustering allows each element to belong with distinct degrees to many clusters, this leads to a flexible representation. Actually, the fuzziness notion is due to the fact that a concept in an ontology is introduced with different attributes and properties allowing it to be assigned to different classes simultaneously.

- Each cluster is introduced with a specific data called medoid which represents the cluster's semantic content. The alignment process is carried over the medoids to determine similar/ correspondent clusters.

- To perform the clustering algorithm, we propose a novel semantic similarity measure. The latter exploits the relational context of the concept to determine the similar concepts.

The alignment process which draws advantage from the ontology hierarchical fuzzy clustering starts by aligning both source and target clusters sets. This alignment step

is performed iteratively over both hierarchical clusters. Starting from the upper level of the source structure, a cluster is selected and compared with target clusters of correspondent level using a semantic similarity technique. Once, the most similar cluster is selected, we move to their correspondent nodes to determine the semantically close ones. The process is repeated until we reach the most similar clusters of the lowest levels. The third step is about aligning entities of the closest clusters (determined in the previous step). For this, the structural and syntactic similarity techniques are computed to discover the aligned entities. Consequently, the problem is reduced from aligning two large ontologies to two correspondent clusters.

In the next section, the characteristics of the proposed approach are illustrated, in Section 3 we introduce the hierarchical fuzzy clustering method; Section 4 describes how the ontology alignment process can explore the hierarchical fuzzy clustering. Section 5 evaluates the performance of the method using the benchmark ontologies of OAEI 2010 and real mammographic ontologies.

2. Ontology Knowledge Mining

2.1. The Fuzzy C-Medoid for ontology's concepts clustering

The Fuzzy c-Medoids FCMdd clustering technique represents a variant of the FCM technique applied to relational data [11]. The use of a fuzzy technique is due to the fact that the ontology concepts are introduced with different attributes and properties so to be assigned to different clusters simultaneously. Likewise, the FCMdd algorithm allows computing the membership degrees of the concepts to the different clusters as well as medoids which represent the representative data of the clusters.

Let $X = \{x_1, \dots, x_n\}$ be a set of n ontology concepts. $d()$ denotes the distance between two concepts of X (defined in section 2.2). $V = \{v_1, \dots, v_c\}$ represents a subset of X with cardinality c(number of clusters); V represents the medoids correspondent to the clusters.

FCMdd is an iterative algorithm which tends to minimize this objective function:

$$J_M(X,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \; d(x_k - v_i)^2 \qquad (1)$$

Where u_{ik} represents the membership degree of concept x_k to the cluster C_i with $\sum_{i=1}^{c} u_{ik} = 1$; m is the fuzziness parameter of the resulting clusters (where $m > 1$), $d()$ is the semantic distance (defined in section 2.2).

Let x_k be a concept of the ontology, v_i and v_j are the medoids which correspond respectively to the clusters C_i and C_j, the membership degree of x_k to C_i is defined as well:

$$u_{ik} = \frac{\left(\frac{1}{d(x_k, v_i)}\right)^{\frac{1}{m-1}}}{\sum_{j=1}^{c}\left(\frac{1}{d(x_k, v_j)}\right)^{\frac{1}{m-1}}} \qquad (2)$$

Once the membership degrees are computed, the algorithm proceeds to determine the medoids v_i for each cluster i. Namely, the medoid of a group of entities is the concept that has the minimal average distance with respect to the others. Formally the medoid of the cluster C, where $v_i, c_j \; \varepsilon \; C$; w.r.t.the semantic distance $d(.)$:

$$v_i = argmin_{v_i \epsilon C}\left(\frac{1}{n}\sum_{j=1}^{n} d(v_i, c_j)\right) \qquad (3)$$

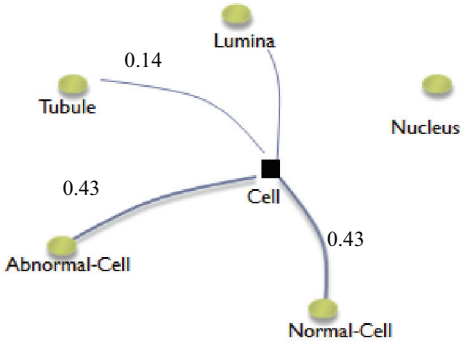

Figure 1: Adherence of the concept 'cell' to the clusters (represented by their medoids)

Actually the medoids designate the concepts minimizing the distance to the other entities of the cluster e.g in the alignment step; those prototypes may intentionally speed-up the task of searching closest clusters.

An example given in Figure 1 denotes the assignment of the concept 'cell' to the different clusters (which are represented through their medoids) where the values represent the membership degree of the concept to the different classes simultaneously.

2.2. Ontology Similarity Measures

The distance (or dissimilarity) based clustering between two ontological elements is a numeric value expressing how close they are. As the clustering quality is highly dependent on the entities constituting the clusters, the used distance affects noticeably the final results. Several similarities for ontologies concepts have been proposed in literature such as the path-based similarity measure. This measure [12] represents a standard form of the path length measure, which is the inverse of the shortest path length between the two terms:

$$\mathrm{sim}(c_1, c_2) = \frac{2N}{N1+N2} \quad (4)$$

Where $N1, N2$ represent respectively the number of arcs between the concepts c_1, c_2 and the root, N is the distance between the lowest concept LCA subsuming c_1, c_2 and the root. This measure has been used for the ontology crisp partitioning in [13] and [14]. It relies on the fact that the deeper the concepts are in the ontological hierarchy, the more they are similar. However, the calculation of the similarity only cumulates the shortest paths together. Another similarity measure has been proposed in [27], it determines the semantic similarity between two concepts based on the information content (IC) of their lowest common ancestor (LCA) node. The information content (IC) gives a measure of how specific and informative a term is. Concepts are considered to be similar if the IC of their LCA is high.

$$\mathbf{sim}(c_1, c_2) = \frac{2\mathbf{IC}(\mathbf{LCA}(c))}{\mathbf{IC}(c_1)+\mathbf{IC}(c_2)} \quad (5)$$

This measure has been used in [15] for the Gene Ontology partitioning. The problem is that methods based on the information content may be inaccurate due to shallow annotations or when ontologies input lack of meaningful information. The above mentioned similarity measures are the most frequent ones used for the ontology clustering. In the following description, we introduce a new semantic similarity based on concept relational context.

In order to cope with these limitations, we propose a new similarity measure based on the *relational context* of a concept. The idea is to define for each concept a relational context that holds the entities to which the concept is related in the ontology. For this, we distinguish two kinds of relationship: First, the subsumption relation that gives information about concepts subsumed by the concept of interest or the concept that subsume it. This kind of relation reflects the elementary structure for a given concept in the ontological hierarchy. Second, the object property relation which reveals the connected concepts and specifies in what sense the object is related to the other object in the ontology. Together, the set of relations describes the semantics of a given concept. Thus, the relational context generates all the concepts closely related to the concept of interest (through the relations mentioned above), in addition to the concept itself. Formally, it is defined as well: Given C the set of concepts in a given ontology, R the set of relations including the subsumption and object property relations, the relational context of a concept $c \in C$ is given by:

$$Cont(c) = \{c_i | (c, c_i) \in R \cup \{c\}\}$$

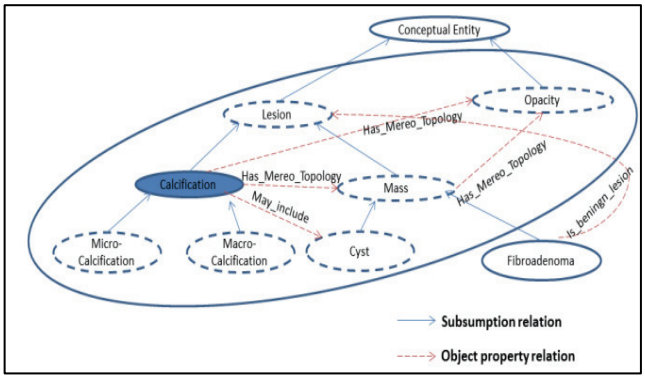

Figure 2: Relational context of the concept 'Calcification'

Figure 2 gives an example of the relational context of the concept 'Calcification' in the mammographic ontology, where we can see that it is related w.r.t the subsumption relation to {Micocalcification, Maco-calcification, Lesion} and w.r.t. to the object-property relation to {Cyst, Mass, Opacity}, then, we can define the relational context of the concept 'Calcification' as {Calcification, Mico-calcification, Macocalcification, Lesion, Cyst, Mass, Opacity}.

Given two concepts c_i and c_j, the distanced (c_i, c_j) based on the relational context is given as well:

$$\mathbf{d}(c_i, c_j) = 1 - \left(2.\frac{|\mathbf{Cont}(c_i) \cap \mathbf{Cont}(c_j)|}{|\mathbf{Cont}(c_i)| + |\mathbf{Cont}(c_j)|}\right) \quad (6)$$

$|Cont(c_i) \cap Cont(c_j)|$ represents the number of common elements between the contexts of c_i and c_j.

$|Cont(c_i)| + |Cont(c_j)|$ represents the sum of both contexts' size, and used for normalization.

2.3. The Fuzzy divisive Algorithm of the ontology's concepts

We present in this section the description of the pre-alignment step based on the ontology knowledge mining. The latter is based on reorganizing the concepts into a hierarchical structure of fuzzy clusters reflecting the knowledge based granularity. The main objective is to group similar concepts hierarchically structured through an iterative process based on a fuzzy divisive clustering algorithm. The process starts with one cluster containing the whole ontology's concepts, then it repeatedly breaks these clusters into more specific and smaller clusters until stopping criterion is satisfied.

At each level, each candidate cluster is verified if it can be further split (e.g. endowed with the least average inner similarity or cohesiveness according to the density measure). In this case, the hierarchy is enlarged by including new fuzzy clusters. Given a set of elements X, the process starts building large clusters. In the initialization of this step, the user selects a number c of clusters so to be as general as possible (c is fixed by a domain expert and it depends on the domain of interest). This step is intended to reflect the user's interests and categorize the domain into groups of targeted thematic, this facilitates knowledge granularity visualization, information retrieval, alignment propagation, etc. Then iteratively the FCMdd algorithm is applied following a top down direction. Each cluster in the previous step is verified whether it can be partitioned into binary fuzzy clusters. Such criterion is important to promise well significant hierarchical clusters. Besides, two parameters are introduced for controlling the loop; the cluster's size and the cluster density. The first parameter is introduced in order to penalize the too many small clusters that could influence the clusters quality. The second parameter is the cluster density or intra-cluster quality, if this latter is less than a predefined threshold α, then this means that elements within the cluster are well correlated and splitting it may not produce the optimal structure of the cluster. The density criteria or intra-cluster quality of the cluster X is defined by:

$$\boldsymbol{\Delta(X) = 2.\left(\frac{\sum_{x_i \varepsilon C} \mu_i \, d(x_i, X)}{\sum_{x_i \varepsilon C} \mu_i}\right)} \qquad (7)$$

In the case where the cluster X of the hierarchical level L of the tree is worth splitting, we apply again the FCMdd Algorithm.

Let X' be a fuzzy sub-cluster of X with membership degree $\mu'_i(x)$ of x, this cluster is added to the hierarchical level $L+1$ as child node of X. The membership degree of x to each new sub cluster of X is defined as the combination of its membership to original cluster X and its membership to X'. This is $\mu'_i(x) * \mu(x)$. In this way, the hierarchical structure states at any point of the iteration a fuzzy partition. Moreover, the concepts of X with low membership degree will have equally low influence in the fuzzy sub-clusters of X. The process iterates till no more clusters are evaluated as worth splitting. It is worth to note that the initialization of the algorithm is realized as well: the farthest two concepts are initialized as medoids of the new fuzzy clusters. In fact of matter, this allows the fast convergence of the algorithm.

The iterative approach, allows automatically determining the optimal number of main clusters. A fuzzy cluster of the hierarchical level L is presented by C_{iL}. Each cluster C_{iL} is represented as a fuzzy subset of concepts:

$$C_{iL} = \{\mu_{iLk} \,/\, x_k, x_k \, \varepsilon X, i = 1, n\}$$

Each fuzzy cluster of a level L is introduced with a medoid v_{iL}(N is the number of sub clusters per level): $V_{iL} = \{v_{1L} \ldots v_{nL}\}$.

3. Ontology Alignment

The ontology knowledge mining process outlined above supplies a great amount of useful information about the levels of granularity of the knowledge base being clustered. The application of this algorithm enables the experts to analyze the deepness and exhaustiveness of the knowledge bases, where domain specific ontologies may be overlapped in the highest levels of the hierarchal structures. This similarity is, then, decreasing while clusters become more and more specified. Likewise, the ontologies alignment process can draw advantage. In this section, we propose an approach for ontologies alignment which exploits the hierarchical fuzzy clustering introduced in previous section. The proposed alignment process is based on two main phases exploring different similarity techniques. The main phases are: *anchor phase* and *derivation phase*.

Algorithm1: Clusters Alignment algorithm

Input: Cand_cluster: Target Cluster
Source_Hierarchy: Source Ontology
Output:
Source_main_cluster
Begin
1: $m_{TC} \leftarrow medoid(CandCluster)$
2: $m_i^L \leftarrow$ *medoid (Source Cluster* i *of level L)*
3: $L \leftarrow$ Root
4: **For** L = Root, Main clusters Level
5: Max=0;
6: CurrCluster$^L \leftarrow m_1^L$;
7: **For** m_i^L in Level L
8: $sim_{semantic}(m_i^L, m_{TC})$// Semantic similarity
9: **If** $sim_{semantic}(m_i^L, m_{TC})$> Max **then**
10: **Begin**
11: CurrCluster$\leftarrow C(m_i^L)$;
12: Max$\leftarrow sim_{semantic}(m_i^L, m_{TC})$;
13: **End**
14: CurrCluster$\leftarrow C(m_i^{L-1})$; //Move to sub-nodes of CurrCluste
15: **Return** CurrCluster;
16: **End**

The Anchor phase consists on aligning the source and target clusters, the main idea is to compare both hierarchical clusters (see Algorithm 1). Once the most similar cluster is retained (line 7 to 12), we re-compare their correspondent sub clusters (line 14) until reaching the lowest level. As the revealed medoids refer to the most representative and descriptive concepts that well characterize the clusters, the similarity is carried over these specified data of the clusters using the semantic similarity based on an external resource. Similar clusters are retained. If the semantic similarity between two medoids is greater than a pre-defined value ω (ω ε[0, 1]), then the process moves to the child nodes. Finally the problem is reduced from aligning two ontologies to aligning couple of clusters where each pair of the similar clusters represents a separate aligning task that is individually solved.

The second phase called the derivation phase is then carried to fully align the elements inside the retained similar clusters. At this step, different similarity measures are applied [17].

3.1. Semantic similarity technique

Semantic based similarities are computed using an external resource. Such similarity is useful when synonyms or semantically closed concepts are used for similar entity in ontologies. The WordNet [16] is an English-language lexical resource that groups words (nouns, verbs, adjectives and adverbs) into sets of synonyms called synsets or term description sets. The synset contains all the terms denoting a given concept. They are linked by semantic relationships such as generalization relationship or specialization relationship. The equation (8) below calculates the synsets similarity value; where A and B designate respectively the synsets of two medoids c_1, c_2.

$$sim_{semantic}(c_1, c_2) = = max(\frac{A \cap B}{A \cup B}) \quad (8)$$

3.2. Syntactic similarity technique

This technique is computed over labels characterizing the couples of entities to be compared. For this, we have used a similarity based Edit-distance which consists on comparing two strings and computing the number of edits (insertions, deletions and substitutions) of required characters to transform one word into another. The syntactic similarity equation of two concepts c_1, c_2 is defined in (9), where $ed(c_1, c_2)$ is the Edit-distance:

$$sim_{syn}(c_1, c_2) = \frac{1}{1+ed(c_1,c_2)} \quad (9)$$

3.3. Structural similarity technique

This similarity measure takes into consideration the position of concepts with respect to their respective taxonomy. It is necessary to check if the concept under consideration is surrounded (descendants and generalizing) by similar concepts in the target ontology.

$$\mathrm{sim}_{\mathrm{Struc}}(c_1, c_2) = \frac{(\mathrm{Sc}(c_1,O1) \cap \mathrm{Sc}(c_2,O2))}{|\mathrm{Sc}(c_1,O1) \cup \mathrm{Sc}(c_2,O2)|} \quad (10)$$

Where $\mathrm{Sc}(c_1, O_1)$denotes the descendants and generalizing of the concept c_1 in the ontologyO_1, and $\mathrm{Sc}(c_2, O_2)$ refers to the descendants and generalizing of the concept c_2 in the ontologyO_2.

4. Evaluation

4.1. Results on benchmark dataset

This section presents the performance evaluation of our method FHCbM applied on the OAEI benchmark data sets and compared with other systems participated in 2010 [18]. We use this version, since the gold standard results of each alignment as well as the performance results of multiple partitioning and non-partitioning systems are available. The OAEI benchmark data sets include a number of ontologies with varied levels of complexities. The base ontology is test-101 considered as the reference/target ontology while the rest represent the source ontologies. The descriptions of these tests are provided in Table 1, they are containing mainly three sets: simple tests (1xx), systematic tests (2xx) and real-life ontologies (3xx).

Table 1. Description of the benchmark data sets.

	Characteristics
101-104	Similar both in entity name and hierarchy structure
201-210	Different linguistic in some levels Similar in hierarchy structure
221-247	Different in hierarchy structure Similar in label description
248-266	Different in both entity names and hierarchy structure
301-304	Real world ontologies conceived by different communities

Most of the alignment systems in comparison accomplished good results for ontologies of the group 1xx in both precision and recall evaluation. This due to the fact that the ontologies of this group have merely similar entity names as well as hierarchy structure. Note that, for ontologies of this benchmark, the first step of the hierarchical fuzzy clustering phase has been performed with (c=2). As there is no linguistic heterogeneity, the system has achieved perfect results in precision and recall as shown in Figure3. For the group 2xx, our system has behaved differently according to the alteration type. For example, in test ontologies 201-202, 206-208, the system has failed in discovering correspondent alignments since the linguistic features of the candidate ontologies are completely modified and the concepts have no names. However, in test ontologies with structural changes, the system is able to discover most of the alignments from the semantic as well as syntactic perspective and both precision and recall values are pretty good. When dealing with synonyms such as ontology 205, the system proved to be effective and ontologies entities were correctly aligned.

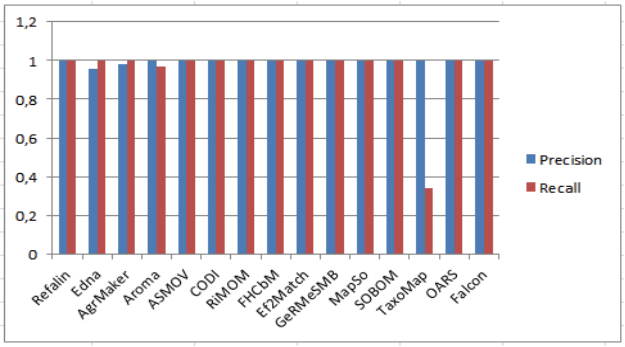

Figure 3: Evaluation results on group 1xx.

For ontologies tests with semantic and structural information are suppressed, it was hard to recognize the correct alignments due to lack of the semantic meanings of classes. As the main step of the alignment process of our system is totally based on the use of the semantic matcher, in the tests of group 2xx the system has not achieved good results of precision and recall regarding other systems in comparison such as RiMOM and Coma as shown in Figure 4.

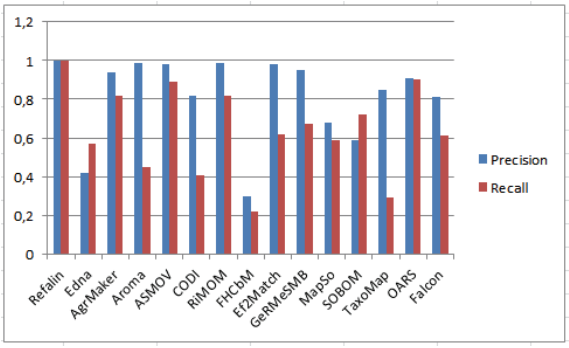

Figure 4: Evaluation results on group 2xx.

In the group 3xx, four real life ontologies of bibliographic references are proposed. Although, the hierarchy structure information is not complete, the results show that our system FHCbM is one of the most effective methods among other systems where the semantic comparability between the two candidate hierarchies in both ontologies in the anchor as well as derivation phases. Figure 5 shows that FHCbM produces good results among other systems with a precision value of 0.92.

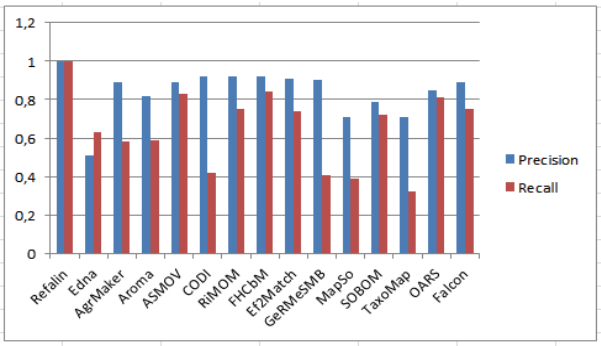

Figure 5: Evaluation results on group 3xx.

4.2. Results on mammographic ontologies

This work is part of CMCU project for Mammographic images analysis for Assistance Diagnostic Breast Cancer, where the main task is to develop a mammographic ontology so to be integrated in support system decision.

Therefore, the proposed approach has been evaluated with experimentations on real world mammographic ontologies which are open source, namely:

-'Breast Cancer Grading Ontology (BCGO)' [19]: The BCGO ontology has been developed in 2009; it contains 364 classes, 156 properties and 164 individuals. It is designed to be application oriented ontology and addresses the problem of semantic gap between high-level semantic concepts and the characteristics of the low-level image.

-'Gimi Mammography ontology' [20]: The Gimi mammography ontology has been developed in 2012; it contains 310 classes and 135 properties, it is used to describe the richness and complexity of the domain and has been implemented with OWL 2, so to be integrated into a learning tool to compare the reviews of trainees with respect to the expert annotations.

The hierarchical fuzzy clustering algorithm has been implemented in Java with the Jena API with setting parameters as well: m = 2. The application gets as input: the ontology to be clustered, a reference file with information about the concepts contexts, and the similarity matrix of concepts. We conducted the evaluations on the ontologies introduced above. The Source ontology is the BCGO ontology where the first hierarchical level handles four categories fixed by radiological experts reflecting the main suitable seeds 'Anatomical_entities', 'Conceptual_entities', 'Descriptors', 'Diagnosis'. We have obtained 5 hierarchical levels: 4 clusters at the top level 17 clusters at the main level. Note that clusters of different size are produced where most specific clusters are deeper with less number of concepts.

As output, a set of fuzzy clusters are generated with estimation of medoids and membership degrees of concepts to the respective clusters.

a. Clustering evaluation:

In order to assess the approach efficiency, the stability of clusters from the similarity perspective is evaluated. For this, the proposed semantic distance has been compared to the 'structural similarity measure'. This measure [12] has been extensively used for ontology partitioning [14] [13]:

$$\mathrm{sim}(c_1, c_2) = \frac{2N}{N1+N2} \qquad (11)$$

Where N_1, N_2 represent the number of edges between concepts c_1, c_2 and the root, N is the distance between the lowest concept subsuming c_1, c_2 and the root.

For the clustering evaluation, as pointed out in several clustering surveys, it is better to use different criteria for clustering. We have used the standard cluster validity measures: Partition coefficient (PC), Partition Entropy (PE) and Purity. Both of PC and PE are based on membership values. The PC indicates the average relative total of membership sharing among pairs of fuzzy subsets [21], the values rang is $[\frac{1}{c}, 1]$ (c is the number of clusters) where a high PC score indicates a better partitioning. The PE reveals the repartition of entities within the clusters [22], the values rang is $[0, log\ c]$, where a low score of PE indicates a better quality of partitioning. Purity of a cluster is the fraction of the highest number of elements shared with other clusters to the total number of elements in the cluster. Correspondent Formulas are given in Table 2. Since, PC and PE work on flat clusters, we have been interested to the lowest level (main clusters) of the hierarchy.

Table 2. Cluster validity measures for fuzzy clustering.

Validity metrics	Characteristics
Partition coefficient (PC)	$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}{\mu_{ij}}^2$
Partition Entropy (PE)	$-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}[\mu_{ij}\ log_2\mu_{ij}]$
Purity	$\frac{1}{\lvert C_i\rvert}max_j(C_i \cap C_j)$

Table 2 illustrates the cluster validity measures for fuzzy clustering where n designates the number of concepts and μ_{ij} is the membership degree of the concept i to cluster j.

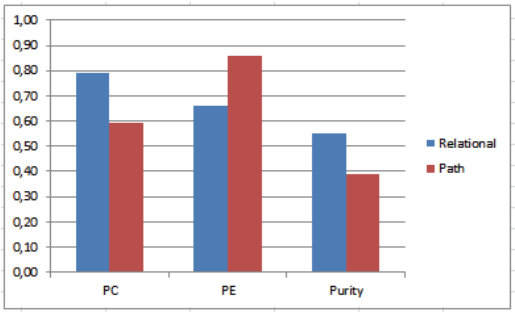

Figure 6: Performance evaluation of the fuzzy clustering based on both ontologies distances

As show in Figure 6, the proposed method has consistently shown better performance with respect to the existing semantic similarity with low PE and high purity and PC. We have noticed also that, with the use of the structural proximity based distance, concepts with weak depth tend to have low membership to different classes. Moreover, medoids designate, generally, concepts with increased depths. Or, this may lead to insignificant representative medoids.

For the ontology hierarchical divisive clustering, a non-comparative evaluation has been realized since to the best of our knowledge no similar methods have been proposed except [23] [24] which are description-language dependent and produce flat clusters. the proposed method in [25] is an instance-based method applied only on populated ontologies and considers both TBox and Abox of the knowledge base clustering.

For the algorithm evaluation, we have assessed each individual cluster as well as analyzed the hierarchy quality, whether sub-clusters of a given class in the hierarchy are well linked. This is by comparing the contents of clusters at each level with the content of corresponding reference clusters (with manual clustering); where we consider concepts with highest membership in each cluster, this evaluation is realized by the means of precision and recall metrics.

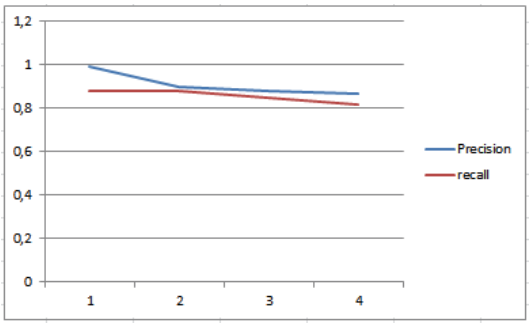

Figure 7: Precision and recall of each hierarchical level for the BCGO

Figure 7 shows the average clustering precision and recall of the proposed algorithm per level, it can be noticed that as clusters are becoming more specific, semantically related concepts remain clustered together; consequently hierarchical linking can place successfully correspondent clusters to the parent cluster.

b. Alignment evaluation:

To evaluate the alignment quality, we adopt 3 standard known metrics widely used in data mining field: Precision, Recall and F-measure. We assume that M designates the set of correspondences discovered between ontological entities by the proposed tool. R is the set of reference correspondences found by the domain expert. These metrics are defined as follows:

- Precision(P): which represents the proportion of true positives among all matching elements found by the method. This allows qualifying the relevance of the alignment method.

-Recall(R): indicates the proportion of true positives among all matching elements in the reference alignment. It quantifies the coverage of the alignment method.

$$P = \frac{|M \cap R|}{|M|} \quad (12) \qquad R = \frac{|M \cap R|}{|R|} \quad (13)$$

-Fmeasure: represents the harmonic mean between precision and recall. It compares the performance of methods by means of single measure.

$$Fmeasure = 2.\frac{\text{Precision.Recall}}{\text{Precision+Recall}} \quad (14)$$

We compare in Figure 8 the performance of our ontology alignment technique with some existing algorithms which are considered among the best performing algorithms FALCON-AO [26] and S-match with respect to a reference alignment which is done manually.

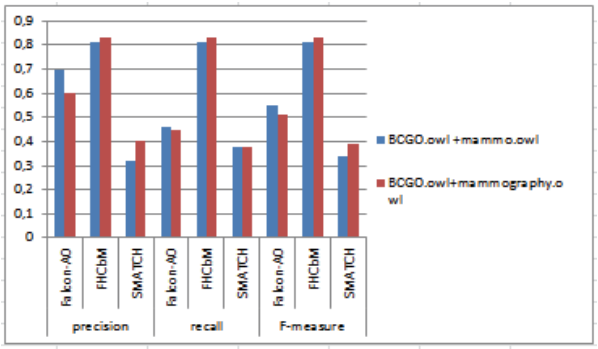

Figure 8: Synthesis of FHCbM, Falcon and SMatch methods- Precision, Recall and F-measure

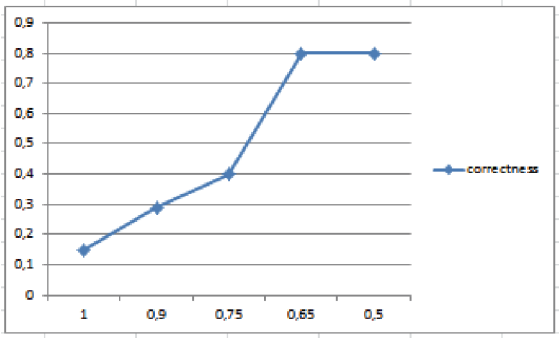

Figure 9: Correctness of alignment results with the variation of ω.

The Falcon-AO is a method that is based on partitioning the ontologies into crisp clusters before aligning the blocks. As regarding the S-Match tool, it is based on non-partitioned strategy; but it uses structural as well as element-based similarity techniques for correspondences determination. We have selected these systems since they are accessible online. The results indicate that our hierarchical fuzzy clustering-based method achieves a slight improvement in alignment quality as compared to the other existing tools. It could identify correct alignment of similar clusters. The reduced search space performs good precision by reducing the total of false positives number. Although the Falcon-AO system adopts ontology partitioning technique to reduce the complexity of the alignment problem, the proposed method is more efficient. As first observation, the use of fuzzy clustering has positively influenced. Specifically, on average, the precision is improved with a variance of 20 percent against the Falcon system 40 percent compared to SMatch system and the recall is enhanced with a variance of 30 percent. This confirms that:

-The use of clustering technique may reduce noticeably the scalability problem by reducing the search space.

-Assigning a concept to several clusters simultaneously increases the chance of finding correct alignments.

Then, we evaluate the anchor phase quality with the variation of the key parameter: the threshold ω of the semantic similarity used in the anchor phase. By using the reference alignments, the correctness of the FHCbM is shown in Figure 9. It can be noticed that, as ω decreases the correctness of the alignment results increases. We can also see when ω>=0.65, the correctness is still larger than 0.8 and the alignment performs pretty good results. Therefore, we have chosen ω=0.65 as the threshold of the semantic similarity in the anchor phase.

We have also been interested in analyzing the subsets found by each matcher (structural, semantic and syntactic) among the resulting alignments. For this, first, we have collected the alignments generated and inspected the subsets discovered by each matcher. The analysis is clarified in Figure 10 where we illustrate how each matcher contributes in the alignment task. Moreover, an important number of alignments are discovered from both or more matchers. Besides, it is well notable that more equivalence alignments come from the syntactic and semantic matchers.

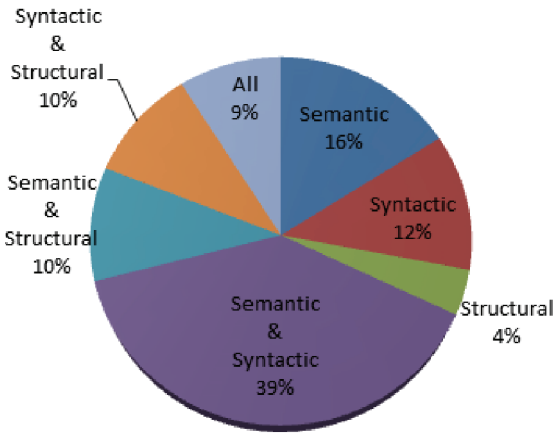

Figure 10: Percentages of the alignments found by the different similarity techniques (among correct alignments)

The semantic matcher proved to be effective to select the appropriate cluster to reason on. The syntactic is the second most reliable matcher next to the semantic one; this means that the syntactic perspective is more trustworthy than from the structural perspective. This is due to the different levels of granularity between large ontologies. Also, medical terms are, in general, standard and typical, that's to say it is rare to find synonyms in medical jargon. Finally, we can conclude that medical ontologies with structural heterogeneity can be successfully tackled because, when this information is suppressed, the syntactic and semantic similarities can well perform in the ontologies alignment.

5. Conclusion

In this paper we have presented the FHCbM, an ontology alignment system using the ontology knowledge mining. The use of the hierarchical fuzzy clustering has proven to be effective in mapping entities. The significance of using clustering techniques was evaluated and compared with a number of existing alignment systems using the benchmark ontology data sets of the OAEI 2010 and real mammographic ontologies. The evaluation results are highly encouraging.

Currently, we are investigating the verification process of FHCbM in order to improve its performance in precision without degrading recall. For this purpose, we are planning to perform the reorganization of ontologies in the pre-alignment step. We also plan to participate in the OAEI campaign in the future. It is worth noting that FHCbM in its current form require the user intervention to initialize the process. Therefore, we are planning to integrate background domain knowledge to automatize the initialization.

References

1 . S. Jan, M. Li, H. Al-Raweshidy, A. Mousavi et M. Qi, «Dealing with uncertain entities in ontology alignment using rough sets,» *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 42, n°16, pp. 1600-1612, 2012.

2. G.Diallo, «An effective method of large scale ontology matching,» *Journal of Biomedical Semantics,* pp. 1-19, 2014.

3. J. Li, J.Beijing, J. Tang, Y. Li et Q. Luo, «RiMOM: A Dynamic Multistrategy Ontology Alignment Framework,» *IEEE transaction on nowledge and Data Engineering,* vol. 21, n° 18, pp. 1218-1232, 2009.

4. W. Hu, Y. Qu et G. Cheng, «Matching large ontologies: A divide-and-conquerapproach,» *Data and knowledge engineering,* vol. 67, pp. 140-160, 2008.

5. S. Massmann, S. Raunich, D. Aumüller, P. Arnold et E. Rahm, «Evolution of the COMA match system,» *Ontology Matching,* vol. 49, June 2011.

6. F. Hamdi, B. Safar, C. Reynaud et H. Zargayouna., «Alignment-based partitioning of large-scale ontologies,» *SCI,* vol. 292, p. 251–269, 2010.

7. A. Algergawy, S.Massmann et E.Rahm, «A Clustering-Based Approach for Large-Scale Ontology Matching,» *Advances in Databases and Information Systems,* pp. 415-428, January 2011.

8. M. Seddiquia et M. Aono, «An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size,» *Web Semantics,* vol. 7, n° 14, pp. 344-356, 2009.

9. W. Hu, Y. Zhao et Y.Qu, «Partition-Based Block Matching of Large Class Hierarchies,» *Proceedings of the First Asian Conference on The Semantic Web,* p. 72–83, 2006.

10. A. Schlicht et H. Stuckenschmidt, «A flexible partitioning tool for large ontologies,» *IEEE/WIC/ACM International Conference on Web Intelligence, WI,* p. 482–488., December 2008.

11. R.E.Stepp et R.S.Michalski, «Conceptual clustering of structured objects: A goal-oriented approach,» *Artificial Intelligence,* vol. 28, n° 11, p. 43–69 , 1986.

12. Z.Wu et M.Palmer, «Verbs semantics and lexical selection,» *Las Cruces, NM, USA,* 1994.

13. H. Hu, Q.Yuzhong et C.Gong, «Matching large ontologies: A divide-and-conquer approach,» *Data & Knowledge Engineering,* vol. 67, pp. 140-160, 2008.

14. M.Karchoudi et S. Yahia, «Large Ontologies Partitioning for Alignment Techniques Scaling,» *Web Information Systems and Technologies,* p. 5, November 2013.

15. T.C.Havens, «Clustering in relational data and ontologies,» Missouri, 2010.

16. G. Miller, «WordNet: a lexical database for English. Communications of the ACM,» vol. 38, n° 111, pp. 39-41, 1995.

17. R.Idoudi, K.S.Ettabaa, K.Hamrouni et B.Solaiman, «An Evidence Based Approach for Multipe Similarity Measures Combining For Ontology Aligning,» *IEEE International Conference on Image Processing Applications and Systems conference (IPAS),* pp. 1-6, November 2014.

18. J. Euzenat, A. Ferrara, C. Meilicke, J. Pane, F. Schar, P. Shvaiko, O. S.-Z. H.Stuckenschmidt, V. Svatek et C. T. d. Santos, «Results of the Ontology Alignment Evaluation Initiative 2010,» *Proceeding of the 5th ISWC Workshop on Ontology Matching (OM-2010),* pp. 1-35, 2010.

19. A. Bulzan, «bioportal,» [En ligne]. Available: http://bioportal.bioontology.org/ontologies/BCGO. [Accès le 26 june 2009].

20. I.oujilov et P.Taylor, «Mammographic Knowledge Representation in Description Logic,» *Springer,* pp. 158-169, August 2012.
21. W.Wang et Y.Zhang, «On fuzzy cluster validity indices,» *Fuzzy Sets and Systems,* vol. 158, n° 119, p. 2095–2117, 14 March 2007.
22. O. M. Jafar et R. Sivakumar, «Hybrid Fuzzy Data Clustering Algorithm Using Different Distance Metrics: A Comparative Study,» *International Journal of Soft Computing and Engineering (IJSCE),* vol. 3, n° 16, pp. 241-248, January 2014.
23. J.-U.Kietz et K. Morik, «A polynomial approach to the constructive induction of structural knowledge,» *Machine Learning,* vol. 14, n° 12, p. 193–218, 1994.
24. N. Fanizzi, L. Iannone, I. Palmisano et G. Semeraro, «Concept formation in expressive description logics,» *J.-F. Boulicaut, et al. (Eds.) Proceedings of the 15th European Conference on Machine Learning ECML, Lecture Notes in Artificial Intelligence,* vol. 3201, p. 99–113, 2004.
25. N. Fanizzi, C.Amato et F. Esposito, «Conceptual Clustering:Concept Formation, Drift and Novelty Detection,» *The Semantic Web: Research and Applications,* vol. 5201, pp. 318-332, 2008.
26. J.Ningsheng, W. Cheng et Q.Yuzhong, «Falcon-AO: Aligning Ontologies with Falcon,» *K-CAP Workshop on Integrating Ontologies,* pp. 85-91, 2005..
27. P.Resnik. (1995). sing Information Content to Evalutate Semantic Similarity in a Taxonomy. International Joint Conference on Artificial IJCAI, 1, 448-453