Clustering Ontology-Based Metadata in the Semantic Web

Alexander Maedche and Valentin Zacharias

FZI Research Center for Information Technologies at the
University of Karlsruhe, Research Group WIM
D-76131 Karlsruhe, Germany
{maedche,zach}@fzi.de
http://www.fzi.de/wim

Abstract. The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. Recently, different applications based on this vision have been designed, e.g. in the fields of knowledge management, community web portals, e-learning, multimedia retrieval, etc. It is obvious that the complex metadata descriptions generated on the basis of pre-defined ontologies serve as perfect input data for machine learning techniques. In this paper we propose an approach for clustering ontology-based metadata. Main contributions of this paper are the definition of a set of similarity measures for comparing ontology-based metadata and an application study using these measures within a hierarchical clustering algorithm.

1 Introduction

The Web in its' current form is an impressive success with a growing number of users and information sources. However, the heavy burden of accessing, extracting, interpretating and maintaining information is left to the human user. Recently, Tim Berners-Lee, the inventor of the WWW, coined the vision of a Semantic Web[1] in which background knowledge on the meaning of Web resources is stored through the use of machine-processable metadata. The Semantic Web should bring structure to the content of Web pages, being an extension of the current Web, in which information is given a well-defined meaning. Recently, different applications based on this Semantic Web vision have been designed, including scenarios such as knowledge management, information integration, community web portals, e-learning, multimedia retrieval, etc. The Semantic Web relies heavily on formal ontologies that provide shared conceptualizations of specific domains and on metadata defined according these ontologies enabling comprehensive and transportable machine understanding.

Our approach relies on a set of similarity measures that allow to compute similarities between ontology-based metadata along different dimensions. The

[1] http://www.w3.org/2001/sw/

T. Elomaa et al. (Eds.): PKDD, LNAI 2431, pp. 348–360, 2002.

similarity measures serve as input to hierarchical clustering algorithm. The similarity measures and the overall clustering approach have been applied on real world data, namely the CIA world fact book[2]. In the context of this empirical evaluation and application study we have obtained promising results.

Organization. Section 2 introduces ontologies and metadata in the context of the Semantic Web. Section 3 focuses on three different similarity measuring dimensions for ontology-based metadata. Section 4 provides insights into our empirical evaluation and application study and the results we obtained when applying our clustering technique on Semantic Web data. Before we conclude and outline the next steps within our work, we give an overview on related work in Section 5.

2 Ontologies and Metadata in the Semantic Web

As introduced earlier the term ”Semantic Web“ encompasses efforts to build a new WWW architecture that enhances content with formal semantics. This will enable automated agents to reason about Web content, and carry out more intelligent tasks on behalf of the user. Figure 1 illustrates the relation between “ontology”, “metadata” and “Web documents”. It depicts a small part of the CIA world fact book ontology. Furthermore, it shows two Web pages, viz. the CIA fact book pages about the country Argentina and the home page of the United Nations, respectively, with semantic annotations given in an XML serialization of RDF-based metadata descriptions[3]. For the country and the organization there are metadata definitions denoted by corresponding uniform resource identifiers (URIs) (HTTP://WWW.CIA.ORG/COUNTRY#AG and HTTP://WWW.UN.ORG#ORG). The URIs are typed with the concepts COUNTRY and ORGANIZATION. In addition, there is a relationship instance between the country and organisation: Argentina ISMEMBEROF United Nations.

In the following we introduce a ontology and metadata model. We here only present the part of our overall model that is actually used within our ontology-based metadata clustering approach[4]. The model that is introduced in the following builds the core backbone for the definition of similarity measures.

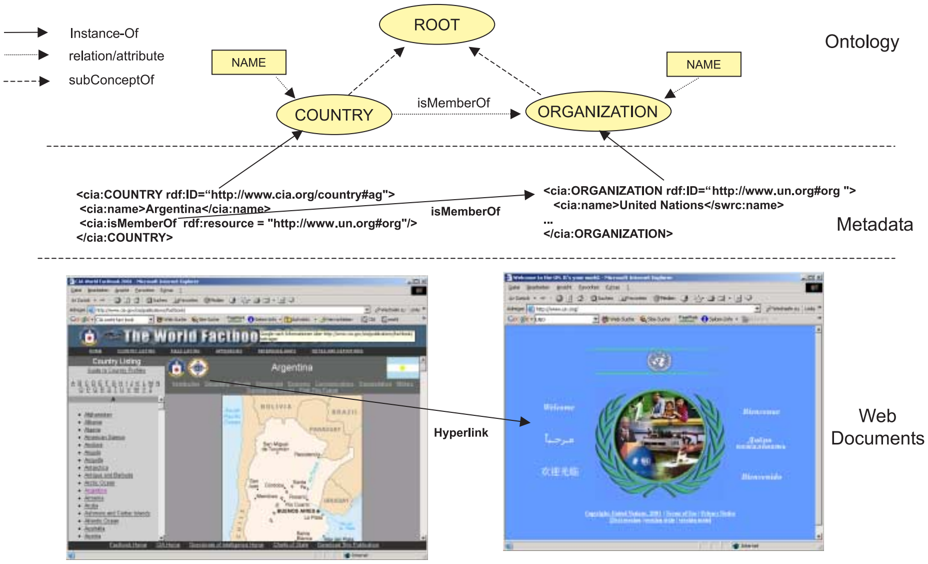
Ontologies. In its classical sense ontology is a philosophical discipline, a branch of philosophy that deals with the nature and the organization of being. In its most prevalent use an ontology refers to an engineering artifact, describing a formal, shared conceptualization of a particular domain of interest [4].

Definition 1 (Ontology Structure). *An ontology structure is a 6-tuple* $\mathcal{O} := \{\mathcal{C}, \mathcal{P}, \mathcal{A}, \mathcal{H}^{\mathcal{C}}, prop, att\}$*, consisting of two disjoint sets* $\mathcal{C}$ *and* $\mathcal{P}$ *whose elements*

[2] http://www.cia.gov/cia/publications/factbook/

[3] The Resource Description Format (RDF) is a W3C Recommendation for metadata representation, http://www.w3c.org/RDF.

[4] A more detailed definition is available in [7].



```
<cia:COUNTRY rdf:ID="http://www.cia.org/country#ag">
 <cia:name>Argentina</cia:name>
 <cia:isMemberOf  rdf:resource = "http://www.un.org#org"/>
</cia:COUNTRY>
```

```
<cia:ORGANIZATION rdf:ID="http://www.un.org#org ">
  <cia:name>United Nations</swrc:name>
...
</cia:ORGANIZATION>
```

Fig. 1. Ontology, metadata and Web documents

are called concepts and relation identifiers, respectively, a ***concept hierarchy*** $\mathcal{H}^{\mathcal{C}}$*:* $\mathcal{H}^{\mathcal{C}}$ *is a directed, transitive relation* $\mathcal{H}^{\mathcal{C}} \subseteq \mathcal{C} \times \mathcal{C}$ *which is also called concept taxonomy.* $\mathcal{H}^{\mathcal{C}}(C_1, C_2)$ *means that* C_1 *is a sub-concept of* C_2*, a* ***function*** $prop : \mathcal{P} \rightarrow \mathcal{C} \times \mathcal{C}$*, that relates concepts non-taxonomically (The function dom:* $\mathcal{P} \rightarrow \mathcal{C}$ *with* $dom(P) := \Pi_1(rel(P))$ *gives the domain of* P*, and range:* $\mathcal{P} \rightarrow \mathcal{C}$ *with* $range(P) := \Pi_2(rel(P)$ *give its range. For* $prop(P) = (C_1, C_2)$ *one may also write* $P(C_1, C_2)$*). A specific kind of relations are attributes* $\mathcal{A}$*. The* ***function*** $att : \mathcal{A} \rightarrow \mathcal{C}$ *relates concepts with literal values (this means* $range(A) := STRING$*)*

Example. Let us consider a short example of an instantiated ontology structure as depicted in Figure 2. Here on the basis of $\mathcal{C} := \{$ COUNTRY, RELIGION, RELIGION$\}$, $\mathcal{P} := \{$BELIEVE, SPEAK, BORDERS$\}$, $\mathcal{A} := \{$POPGRW$\}$ the relations BELIEVE(COUNTRY, RELIGION), SPEAK(COUNTRY, LANGUAGE), BORDERS(COUNTRY, COUNTRY) with its domain/range restrictions and the attribute POPGRW(COUNTRY) are defined.

Ontology-Based Metadata. We consider the term metadata as synonym to instances of ontologies and define a so-called metadata structure as following:

Definition 2 (Metadata Structure). *A metadata structure is a 6-tupel* $\mathcal{MD} := \{\mathcal{O}, \mathcal{I}, \mathcal{L}, inst, instr, instl\}$*, that consists of an ontology* $\mathcal{O}$*, a set* $\mathcal{I}$ *whose elements are called instance identifiers (correspondingly C, P and I are disjoint), a set of literal values L, a function* $inst : \mathcal{C} \rightarrow 2^{\mathcal{I}}$ *called* ***concept instantiation*** *(For* $inst(C) = I$ *one may also write* $C(I)$*), and a function* *instr* : $\mathcal{P} \rightarrow 2^{\mathcal{I}\times\mathcal{I}}$ *called* ***relation instantiation*** *(For* $inst(P) = \{I_1, I_2\}$ *one may also write* $P(I_1, I_2)$*). The* ***attribute instantiation*** *is described via the function* $instl : \mathcal{P} \rightarrow 2^{\mathcal{I}\times\mathcal{L}}$ *relates instances with literal values.*

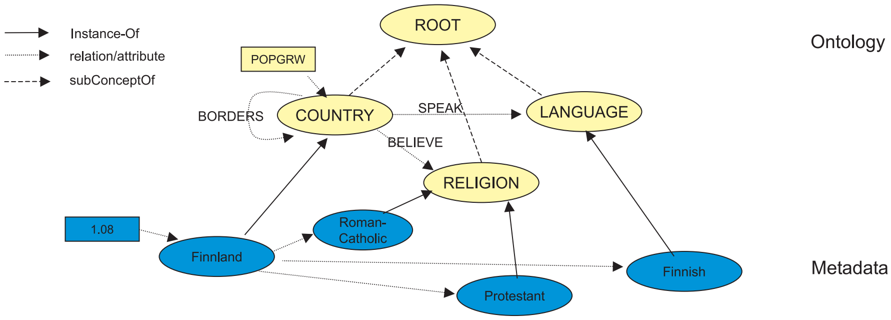


Fig. 2. Example ontology and metadata

Example. Here, the following metadata statements according to the ontology are defined. Let $\mathcal{I}$:= {Finnland, Roman-Catholic, Protestant, Finnish}. *inst* is applied as follows: *inst*(Finnland) = COUNTRY, *inst*(Roman-Catholic) = RELIGION, *inst*(Protestant) = RELIGION, *inst*(Finnish) = LANGUAGE. Furthermore, we define relations between the instances and an attribute for the country instance. This is done as follows: We define BELIEVE(Finnland, Roman-Catholic), BELIEVE(Finnland, Protestant), SPEAK(Finnland, Finnish) and POPGRW(Finnland, "1.08″).

3 Measuring Similarity on Ontology-Based Metadata

As mentioned earlier, clustering of objects requires some kind of similarity measure that is computed between the objects. In our specific case the objects are described via ontology-based metadata that serve as input for measuring similarities. Our approach is based on similarities using the instantiated ontology structure and the instantiated metadata structure as introduced earlier in parallel. Within the overall similarity computation approach, we distinguish the following three dimensions:

- **Taxonomy similarity:** Computes the similarity between two instances on the basis of their corresponding concepts and their position in $\mathcal{H}^{\mathcal{C}}$.
- **Relation similarity:** Compute the similarity between two instances on the basis of their relations to other objects.
- **Attribute similarity:** Computes the similarity between two instances on the basis of their attributes and attribute values.

Taxonomy Similarity. The taxonomic similarity computed between metadata instances relies on the concepts with their position in the concept taxonomy $\mathcal{H}^{\mathcal{C}}$. The so-called upwards cotopy (SC) [7] is the underlying measure to compute the semantic distance in a concept hierarchy.

Definition 3 (Upwards Cotopy (UC)).

$$UC(C_i, \mathcal{H}^{\mathcal{C}}) := \{C_j \in \mathcal{C} | \mathcal{H}^{\mathcal{C}}(C_i, C_j) \vee C_j = C_i\}.$$

The semantic characteristics of $\mathcal{H}^{\mathcal{C}}$ are utilized: The attention is restricted to super-concepts of a given concept C_i and the reflexive relationship of C_i to itself. Based on the definition of the upwards cotopy (UC) the concept match (CM) is then defined:

Definition 4 (Concept Match).

$$CM(C_1, C_2 := \frac{|(UC(C_1, \mathcal{H}^{\mathcal{C}}) \cap (UC(C_2, \mathcal{H}^{\mathcal{C}}))|}{|(UC(C_1, \mathcal{H}^{\mathcal{C}})) \cup (UC(C_2, \mathcal{H}^{\mathcal{C}})|}.$$

Example. Figure 3 depicts the example scenario for computing CM graphically. The upwards cotopy $\text{UC}(\text{CHRISTIANISM}, \mathcal{H}^{\mathcal{C}})$ is given by $(\text{UC}((\{\text{CHISTIANISM}\}), \mathcal{H}^{\mathcal{C}})) = \{\text{CHRISTIANISM}, \text{RELIGION}, \text{ROOT}\}$. The upwards cotopy $\text{UC}((\{\text{MUSLIM}\}), \mathcal{H}^{\mathcal{C}})$ is computed by $\text{UC}((\{\text{MUSLIM}\}), \mathcal{H}^{\mathcal{C}}) = \{\text{MUSLIM}, \text{RELIGION}, \text{ROOT}\}$. Based on the upwards cotopy one can compute the concept match CM between two given specific concepts. The concept match CM between MUSLIM and CHRISTIANISM is given as $\frac{1}{2}$.

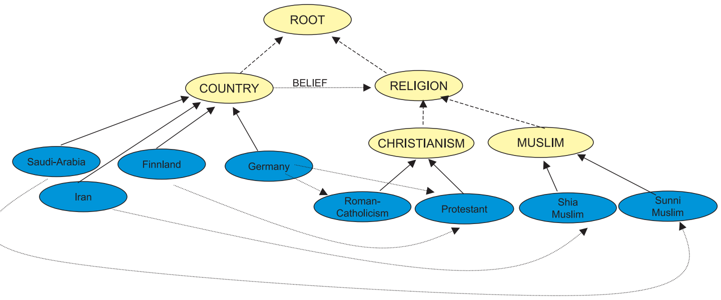


Fig. 3. Example for computing similarities

Definition 5 (Taxonomy Similarity).

$$TS(I_1, I_2) = \begin{cases} 1 \quad if\ I_1 = I_2 \\ \frac{CM(C(I_1), C(I_2))}{2} & otherwise \end{cases}$$

The taxonomy similarity between SHIA MUSLIM to PROTESTANT results in $\frac{1}{4}$.

Relation similarity. Our algorithm is based on the assumption that if two instances have the same relation to a third instance, they are more likely similar than two instances that have relations to totally different instances. Thus, the similarity of two instances depends on the similarity of the instances they have relations to. The similarity of the referred instances is once again calculated using taxonomic similarity. For example, assuming we are given two concepts COUNTRY and RELIGION and a relation BELIEVE(COUNTRY, RELIGION). The algorithm will infer that specific countries believing in catholizism and protestantism are more similar than either of these two compared to hinduism because more countries have both catholics and protestants than a combination of either of these and hindis.

After this overview, let's get to the nitty gritty of really defining the similarity on relations. We are comparing two instances I_1 and I_2, $I_1, I_2 \in \mathcal{I}$. From the definition of the ontology we know that there is a set of relations P_1 that allow instance I_1 either as domain, as range or both (Likewise there is a set P_2 for I_2). Only the intersection $P_{\mathrm{CO}} = P_1 \cap P_2$ will be of interest for relation similarity because differences between P_1 and P_2 are determined by the taxonomic relations, which are already taken into account by the taxonomic similarity. The set P_{CO} of relations is differentiated between relations allowing I_1 and I_2 as range - $P_{\mathrm{CO-I}}$, and those that allow I_1 and I_2 as domain - $P_{\mathrm{CO-O}}$.

Definition 6 (Incoming $P_{\mathbf{CO-I}}$ and Outgoing $P_{\mathbf{CO-O}}$ Relations). *Given $\mathcal{O} := \{\mathcal{C}, \mathcal{P}, \mathcal{A}, \mathcal{H}^{\mathcal{C},\mathcal{P}}, prop, att\}$ and instances I_1 and I_2 let:*

$$H^{trans} := \left\{(a,b) : (\exists a_1 ... a_n \in C : H^C(a, a_1) ... H^C(a_n, b))\right\}$$

$$P_{co-Ii}(I_i) := \{R : R \in \mathcal{P} \wedge ((C(I_i), range(R)) \in H^{trans})\}$$
$$P_{co-Oi}(I_i) := \{R : R \in \mathcal{P} \wedge ((C(I_i), domain(R)) \in H^{trans})\}$$
$$P_{co-I}(I_i, I_j) := P_{co-Ii}(I_i) \cap P_{co-I}(I_j)$$
$$P_{co-O}(I_i, I_j) := P_{co-Oi}(I_i) \cap P_{co-O}(I_j)$$

In the following we will only look at $P_{\mathrm{CO-O}}$, but everything applies to $P_{\mathrm{CO-I}}$ as well. Before we continue we have to note an interesting aspect: For a given ontology with a relation P_x there is a minimum similarity greater than zero between any two instances that are source or target of an instance relation - $\mathrm{MinSim}_{s(P_x)}$ and $\mathrm{MinSim}_{t(P_x)}$[5]. Ignoring this will increase the similarity of two instances with relations to the most different instances when compared to two instances that simply don't define this relation. This is especially troublesome when dealing with missing values. For each relation $P_n \in P_{\mathrm{CO-O}}$ and each instance I_i there exists a set of instance relations $P_n(I_i, I_x)$. We will call the set of instances I_x the associated instances A_s.

Definition 7 (Associated Instances).

$$A_s(P, I) := \{I_x : I_x \in \mathcal{I} \wedge P(I, I_x)\}$$

[5] Range and domain specify a concept and any two instances of this concept or one of its sub-concepts will have a taxonomic similarity bigger than zero

The task of comparing the instances I_1 and I_2 with respect to relation P_n boils down to comparing $A_s(P_n, I_1)$ with $A_s(P_n, I_2)$. This is done as follows:

Definition 8 (Similarity for One Relation).

$$OR(I_1, I_2, P) = \begin{cases} \mathit{MinSim}_{t(P)} & \mathit{if}\ A_s(P, I_1) = \emptyset \vee A_s(P, I_2) = \emptyset \\ \left(\frac{\sum_{(a \in A_s(P,I_1))} \max\{sim(a,b) | b \in A_s(P,I_2)\}}{|A_s(P,I_1)|}\right) & \mathit{if}\ |A_s(P, I_1)| \geq |A_s(P, I_2)| \\ \left(\frac{\sum_{(a \in A_s(P,I_2))} \max\{sim(a,b) | b \in A_s(P,I_1)\}}{|A_s(P,I_2)|}\right) & \mathit{otherwise} \end{cases}$$

Finally, the results for all $P_n \in P_{\text{co–O}}$ and $P_n \in P_{\text{co–I}}$ are combined by calculating their arithmetic mean.

Definition 9 (Relational Similarity).

$$RS(I_1, I_2) := \frac{\sum_{p \in P_{co-I}} OR(I_1, I_2, p) + \sum_{p \in P_{co-O}} OR(I_1, I_2, p)}{|P_{co-I}| + |P_{co-O}|}$$

The last problem that remains is the recursive nature of process of calculating similarities that may lead to infinite cycles, but it can be easily solved by imposing a maximum depth for the recursion. After reaching this maximum depth the arithmetic mean of taxonomic and attribute similarity is returned.

Example. Assuming based on Figure 3 we compare Finnland and Germany, we see that the set of common relations only contains the belief relation. As the next step we compare the sets of instances associated with Germany and Finnland through the belief relation - that's {Roman-Catholicism, Protestant} for Germany and Protestant for Finnland. The similarity function for Protestant compared with Protestant returns one because they are equal, but the similarity of Protestant compared with Roman-Catholicsm once again depends on their relational similarity. If we we assume the the maximum depth of recursion is set to one, the relational similarity between Roman-Catholicsm and Protestant is 0.5[6]. So finally the relational similarity between Finnland and Germany in this example is 0.75.

Attribute Similarity. Attribute similarity focuses on similar attribute values to determine the similarity between two instances. As attributes are very similar to relations[7], most of what is said for relations also applies here.

Definition 10 (Compared Attributes for Two Instances).

$$P_A i(I_i) := \{A : A \in \mathcal{A}\}$$

$$P_A(I_i, I_j) := P_A i(I_i) \cap P_A i(I_j)$$

[6] The set of associated instances for Protestant contains Finnland and Germany, the set for Roman-Catholicism just Germany.

[7] In RDF attributes are actually relations with a range of literal.

Definition 11 (Attribute Values).

$$A_s(A, I_i) := \{L_x : L_x \in \mathcal{L} \wedge A(I_i, L_x)\}$$

Only the members of the sets A_s defined earlier are not instances but literals and we need a new similarity method to compare literals. Because attributes can be names, date of birth, population of a country, income etc. comparing them in a senseful way is very difficult. We decided to try to parse the attribute values as a known data type (so far only date or number)[8] and to do the comparison on the parsed values. If it's not possible to parse all values of a specific attribute, we ignore this attribute. But even if numbers are compared, translating a numeric difference to a similarity value $[0, 1]$ can be difficult. For example comparing the attribute population of a country a difference of 4 should yield a similarity value very close to 1, but comparing the attribute "average number of children per woman" the same numeric difference value should result in a similarity value close to 0. To take this into account, we first find the maximum difference between values of this attribute and then calculate the the similarity as $1 - (\text{Difference}/\max\text{Difference})$.

Definition 12 (Literal Similarity).

$$slsim(\mathcal{A}, \mathcal{A}) \rightarrow [0, 1]$$

$$mlsim := \max\{slsim(A_1, A_2) : A_1 \in \mathcal{A} \wedge A_2 \in \mathcal{A}\}$$

$$lsim(A_i, A_j, A) := \frac{slsim(A_i, A_j)}{mlsim(A)}$$

And last but not least, unlike for relations the minimal similarity when comparing attributes is always zero.

Definition 13 (Similarity for One Attribute).

$$OA(I_1, I_2, A) := \begin{cases} 0 & \mathit{if}\ A_s(A, I_1) = \emptyset \vee A_s(A, I_2) = \emptyset \\ \left(\frac{\sum_{(a \in A_s(A, I_1))} \max\{lsim(a,b,A) | b \in A_s(A, I_2)\}}{|A_s(A, I_1)|}\right) & \mathit{if}\ |A_s(A, I_1)| \geq |A_s(A, I_2)| \\ \left(\frac{\sum_{(a \in A_s(A, I_2))} \max\{lsim(a,b,A) | b \in A_s(A, I_1)\}}{|A_s(A, I_2)|}\right) & \mathit{otherwise} \end{cases}$$

Definition 14 (Attribute Similarity).

$$AS(I_1, I_2) := \frac{\sum_{a \in P_{A}(I_1, I_2)} OA(I_1, I_2, a)}{|P_{A(I_1, I_2)}|}$$

[8] For simple string data types one may use a notion of string similarity: The *edit distance* formulated by Levenshtein [6] is a well-established method for weighting the difference between two strings. It measures the minimum number of token insertions, deletions, and substitutions required to transform one string into another using a dynamic programming algorithm. For example, the edit distance, ed, between the two lexical entries "TopHotel" and "Top_Hotel" equals 1, ed("TopHotel", "Top_Hotel") = 1, because one insertion operation changes the string "TopHotel" into "Top_Hotel".

Combined Measure. The combined measure uses the three dimensions introduced above in a common measure. This done by calculating the weighted arithmetic mean of attribute, relation and semantic similarity.

Definition 15 (Similarity Measure).

$$sim(I_i, I_j) := \frac{t \times TS(I_i, I_j) + r \times RS(I_i, I_j) + a \times AS(I_i, I_j)}{t + r + a}$$

The weights may be adjusted according to the given data set the measures should be applied, e.g. within our empirical evaluation we used a weight of 2 for relation similarity, because most of the overall information of the ontology and the associated metadata was contained in the relations.

Hierarchical Clustering. Based on the similarity measures introduced above we may now apply a clustering technique. Hierarchical clustering algorithms are preferable for concept-based learning. They produce hierarchies of clusters, and therefore contain more information than non-hierarchical algorithms. [8] describes the bottom-up algorithm we use within our approach. It starts with a separate cluster for each object. In each step, the two most similar clusters are are determined, and merged into a new cluster. The algorithm terminates when one large cluster containing all objects has been formed.

4 Empirical Evaluation

We have empirically evaluated our approach for clustering ontology-based metadata based on the different similarity measures and the clustering algorithm introduced above. We used the well-known CIA world fact book data set as input[9] available in the form of a MONDIAL database[10]. Due to a lack of currently available ontology-based metadata on the Web, we converted a subset of MONDIAL in RDF and modeled a corresponding RDF-Schema for the databases (on the basis of the ER model also provided by MONDIAL). Our subset of the MONDIAL database contained the concepts COUNTRY, LANGUAGE, ETHNIC-GROUP, RELIGION and CONTINENT. Relations contained where

- SPEAK(COUNTRY,LANGUAGE),
- BELONG(COUNTRY, ETHNIC-GROUP),
- BELIEVE(COUNTRY,RELIGION),
- BORDERS(COUNTRY,COUNTRY) and
- ENCOMPASSES(COUNTRY,CONTINENT).

We also converted the attributes infant mortality and population growth of the concept COUNTRY. As there is no pre-classification of countries, we decided

[9] http://www.cia.gov/cia/publications/factbook/

[10] http://www.informatik.uni-freiburg.de/~may/Mondial/

to empirically evaluate the cluster against the country clusters we know and use in our daily live (like european countries, scandinavian countries, arabic countries etc). Sadly there is no further taxonomic information for the concepts RELIGION, ETHNIC–GROUP or LANGUAGE available within the data set. For our experiments we used the already introduced bottom-up clustering algorithm with a single linkage computation strategy using cosine measure.

Using only relation similarity. Using only the relations of countries for measuring similarities we got clusters resembling many real world country clusters, like the european countries, the former soviet republics in the caucasus or such small cluster like {AUSTRIA, GERMANY}. A particular interesting example is the cluster of scandinavian countries depicted in Figure 4 because our data nowhere contains a value like ”scandinavian language” or a ethnic group ”scandinavian”.[11] Figure 5 shows another interesting cluster of countries that we know as the Middle East[12]. The politically interested reader will immediately recognize that Israel is missing. This can be easily explained by observing that Israel, while geographically in the middle east is in terms of language, religion and ethnic group a very different country. More troublesome is that Oman is missing too and this can be only explained by turning to the data set used to calculate the similarities, where we see that Oman is missing many values, for example any relation to language or ethnic group.

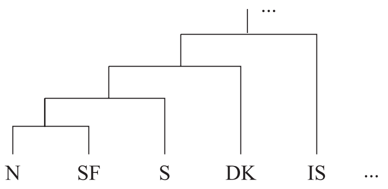


Fig. 4. Example clustering result – scandinavian countries

Using only attribute similarity. When using only attributes of countries for measuring similarities we had to restrict the clustering to infant mortality and population growth. As infant mortality and population growth are good indicators for wealth of a country, we got cluster like industrialized countries or very poor countries.

[11] The meaning of the acronyms in the picture is: N:Norway, SF: Finnland, S: Sweden, DK: Denmark and IS:Island.

[12] The meaning of the acronyms used in the picture is: Q:Quatar, KWT: Kuwait, UAE: United Arab Emirates, SA: Saudi Arabia, JOR: Jordan, RL: Lebanon, IRQ: Iraq, SYR: Syria, YE, Yemen.

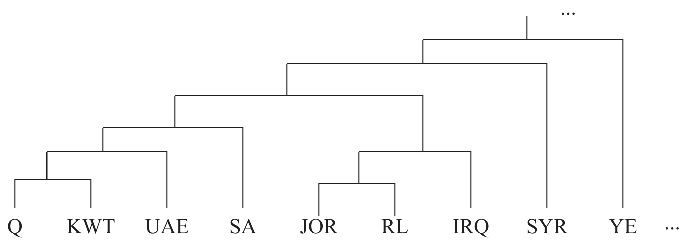


Fig. 5. Example clustering result – middle east

Combining relation and attribute similarity. At first surprisingly the clusters generated with the combination of attribute and relation similarity closely resemble the clusters generated only with relation similarity. But after checking the attribute values of the countries it actually increased our confidence in the algorithm, because countries that are geographically close together, and are similar in terms of ethnic group, religion and language are almost always also similar in terms of population growth and infant mortality. In the few cases where this was not the case the countries where rated far apart, for example Saudi Arabia and Iraq lost it's position in the core middle east cluster depicted because of their high infant mortality[13].

Summarization of results. Due to the lack of pre-classified countries and due to the subjectivity of clustering in general, we had to restrict our evaluation procedure to an empirical evaluation of the cluster we obtained against the country clusters we know and use in our daily live. It has been seen that using our attribute and relation similarity measures combined with a hierarchical clustering algorithm results in reasonable clusters of countries taking into account the very different aspects a country may be described and classified.

5 Related Work

One work closely related to ours was done by Bisson [1]. In [1] it is argued that object-based representation systems should use the notion of similarity instead of the subsumption criterion for classification and categorization. The similarity between attributes is obtained by calculated the similarity between the values for common attributes (taking upper and lower bound for this attribute into account) and combining them. For a symmetrical similarity measure they are combined by dividing the weighted sum of the similarity values for the common attributes by the weights of all attribute that occur in one of the compared

[13] It may be surprising for such a rich country, but according to the CIA world fact book the infant mortality rate in Saudi Arabia (51 death per 1000 live born children) much closer resembles that of sanctioned Iraq (60) than that of much poorer countries like Syria (33) or Lebanon (28)

individuals. For a asymmetrical similarity measure the sum is divided using just the weights for the attributes that occur in the first argument individual, thereby allowing to calculate the degree of inclusion between first and second argument. The similarity for relations is calculated by using the similarity of the individuals that are connected through this relations. The resulting similarity measures are then again combined in the above described symmetrical or asymmetrical way. Compared to the algorithm proposed here the approach proposed by Bisson does not take ontological backgound knowledge into account.

Similar to our approach a distance-based clustering is introduced in [3] that used RIBL (Relational Instance-Based Learning) for distance computations. RIBL as introduced in [5] is an adaption of a propositional instance-based learner to a first order representation. It uses distance weighted k-nearest neighbor learning to classify test cases. In Order to calculate the distance between examples RIBL computes for each example a conjunction of literals describing the objects that are represented by the arguments of the example fact. Given an example fact RIBL first collects all facts from the knowledge base containing at least one of the arguments also contained in the example fact. Depending on a parameter set by the user, the system may then continue to collect all facts that contain at least one of the arguments contained in the earlier selected facts (this goes on until a specified depth is reached). After selecting these facts the algorithm then goes on to calculate the similarity between the examples in a manner similar to the one used by Bisson or described in this paper: The similarity of the objects depends on the similarity of their attribute values and on the similarity of the objects related to them. The calculation of the similarity value is augmented by predicate and attribute weight estimation based on classification feedback[14]. But like Bissons approach RIBL does not use ontological background knowledge[15].

In the context of Semantic Web research, an approach for clustering RDF statements to obtain and refine an ontology has been introduced by [2]. The authors present a method for learning concept hierarchies by systematically generating the most specific generalization of all possible sets of resources - in essence building a subsumption hierarchy using both the intension and extension of newly formed concepts. If an ontology is already present, its information is used to find generalizations - for example generalizing "type of Max is Cat" and "type of Moritz is Dog" to "type of Max,Moritz is Mammal". Unlike the authors of [2] we deliberately chose to use a distance and not a subsumption based clustering because - as for example [2] points out - subsumption based criteria are not

[14] Weight estimation was not used in [3]

[15] It may seem obvious that it is possible to include ontological background information as facts in the knowledge base, but the results would not be comparable to our approach. Assuming we are comparing u1, u2 and have the facts instance_of(u1,c1), instance_of(u2,c2). Comparing u1 and u2 with respect to instance_of would lead to comparing c1 and c2 which in turn lets the algorithm select all facts containing c1 and c2 - containing all instances of c1 and c2 and their description. Assuming a single root concept and a high depth parameter sooner or later all facts will be selected - resulting not only in a long runtime but also in a very low impact of the taxonomic relations

well equipped to deal with incomplete or incoherent information (something we expect to be very common within the Semantic Web).

6 Conclusion

In this paper we have presented an approach towards mining Semantic Web data, focusing on clustering objects described by ontology-based metadata. Our method has been empirically evaluated on the basis of the CIA world fact book data set that was easily to convert into ontology-based metadata. The results have shown that our clustering method is able to detect commonly known clusters of countries like scandinavian countries or middle east countries.

In the future much work remains to be done. Our empirical evaluation could not be formalized due to the lack of available pre-classifications. The actual problem is that there are no ontological background knowledge. Therefore, we will model country clusters within the CIA world fact book ontology and experiment to which degree the algorithm is able to discover these country clusters. These data set may serve as a future reference data set when experimenting with our Semantic Web mining techniques.

Acknowledgments

The research presented in this paper has been partially funded by Daimler-Chrysler AG, Woerth in the HRMore project. We thank Steffen Staab for providing useful input for defining the taxonomic similarity measure. Furthermore, we thank our student Peter Horn who did the implementation work for our empirical evaluation study.

References

1. G. Bisson. Why and how to define a similarity measure for object based representation systems, 1995. 358
2. A. Delteil, C. Faron-Zucker, and R. Dieng. Learning ontologies from RDF annotations. In A. Maedche, S. Staab, C. Nedellec, and E. Hovy, editors, *Proceedings of IJCAI-01 Workshop on Ontology Learning OL-2001, Seattle, August 2001*, Menlo Park, 2001. AAAI Press. 359
3. W. Emde and D. Wettschereck. Relational instance-based learning. Proceedings of the 13th International Conference on Machine Learning, 1996, 1996. 359
4. T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 6(2):199–221, 1993. 349
5. M. Kirsten and S. Wrobel. Relational distance-based clustering. pages 261–270. Proceedings of ILP-98, LNAI 1449, Springer, 1998, 1998. 359
6. I. V. Levenshtein. Binary Codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8):707–710, 1966. 355
7. A. Maedche, S. Staab, N. Stojanovic, R. Studer, and Y. Sure. *SEmantic PortAL – The SEAL approach.* to appear in: Creating the Semantic Web. D. Fensel et al., MIT Press, MA, Cambridge, 2001. 349, 352
8. C. D. Manning and H. Schuetze. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, Massachusetts, 1999. 356