TIN2009-11005
DAMASK

Data-Mining Algorithms with Semantic Knowledge

PROYECTO DE INVESTIGACIÓN
PROGRAMA NACIONAL DE INVESTIGACIÓN FUNDAMENTAL,
PLAN NACIONAL DE I+D+i 2008-2011
ÁREA TEMÁTICA DE GESTIÓN: Tecnologías informáticas

Internal project report

Subtask T-3.4

Distance Measures for Heterogeneous Values

Authored by
Ferran Mata, Universitat Rovira i Virgili
Aïda Valls, Universitat Rovira i Virgili

Document information

project name:	DAMASK	
Project reference:	TIN2009-11005	
type of document:	Internal Report	
file name:	Report T3.4 Distance	
version:	1.0	
authored by:	F. Mata, A. Valls	15/04/2012
co-authored by		
released by:	A.Valls	01/05/2012
approved by:	Co-ordinator	Antonio Moreno

Document history

version	date	reason of modification
1.0	15.April.2012	Definition of the distance functions to be used in the recommender system.
1.1	01.May.2012	Revised document.

Table of Contents

1 Introduction

In Deliverable D3 the dissimilarity and distance functions for numerical and categorical data were presented. In Deliverable D4 the case of semantic similarity was discussed. Also in D4 a compatibility measure was presented to combine different types of attributes into a single measure. It permits the combination of the contribution of numerical, nominal and semantic features into a global function (Batet, 2010).

In case of not having weights for the different attributes, according to the principles of compatibility measures proposed by Anderberg (Anderberg, 1973), the contribution of a single feature to the final distance can be set up depending on its type and it can be computed per blocks, regarding the types of the considered variables. This expression (Eq. 1) permits to associate a weight to each component, giving different importance to numerical (N), categorical (C) and semantic attributes (S).

$$d(i,i') = \alpha \sum_{k \in N} d_k^N + \frac{\beta}{n_C^2} \sum_{k \in C} d_k^C(i,i') + \frac{\gamma}{n_S^2} \sum_{k \in S} d_k^S(i,i') \quad (1)$$

In case of knowing the weight that the user wants to give to each type of attribute, the three components in Eq. 1 can be weighted by the user. The set of weights will fulfil that $w_N + w_C + w_S = 1$.

$$d(i,i') = w_N \sum_{k \in N} d_k^N + w_C \sum_{k \in C} d_k^C(i,i') + w_S \sum_{k \in S} d_k^S(i,i') \quad (2)$$

The quadratic form of the distance $d^2(i,i')$ is required in some clustering methods, such as the Ward criterion that was tested in some previous works (Batet, 2010; Batet, Valls, & Gibert, 2011). Since in the prototype we are going to apply the k-means algorithm, we do not need a quadratic form, as given in Eq. 2.

In the following sections we will give the details about the distance calculation for each type of attribute.

2 Distance calculation

2.1 Numerical attributes

In the CITIES data matrix, the numerical attributes are two:

- Population – Numerical
- Elevation – Numerical

The distance will be calculated with the Euclidean distance, as proposed in Deliverable D4. To allow the user to give different overall importance to each feature, we have implemented the weighted Euclidean distance for the attributes in the set *N*. As usual, we consider that $\sum_{k=1}^{|N!} w_k = 1$.

$$d_k^N(i,i') = \sqrt[2]{\sum_{k=1}^{|N|} w_k (x_{ik} - x_{i'k})^2} \quad (3)$$

It is worth to note that the values x_{ij} are previously normalized in the range [0, 1] using Eq. 4.

$$x_{norm} = \frac{x - min}{\max - min} \quad (4)$$

Due to the fact that the frequency distribution is not uniform but has a high peak on the low values (see Internal Report T3-2), we have taken as maximum value the one at the percentile 85%. Consequently, for the Population attribute, the maximum is fixed at 4,000,000 and cities with highest concentrations of people will receive a normalized value of 1. For the Elevation attribute, the maximum is fixed at 250 meters.

The following tables show some examples of the results obtained.

Table 1. Distances between some cities according to the attribute Population

	Aberdeen	Abu_Dhabi	Agra	Amsterdam	Antwerp	Atlanta	Bahrain	Bangkok	Barcelona	Bath,_Some	Beijing
Aberdeen	-	0,104	0,310	0,139	0,069	0,059	0,138	1,000	0,358	0,023	1,000
Abu_Dhabi	0,104	-	0,206	0,034	0,036	0,046	0,033	1,000	0,253	0,127	1,000
Agra	0,310	0,206	-	0,171	0,241	0,251	0,172	0,914	0,048	0,333	1,000
Amsterdam	0,139	0,034	0,171	-	0,070	0,080	0,001	1,000	0,219	0,161	1,000
Antwerp	0,069	0,036	0,241	0,070	-	0,010	0,069	1,000	0,289	0,091	1,000
Atlanta	0,059	0,046	0,251	0,080	0,010	-	0,079	1,000	0,299	0,081	1,000
Bahrain	0,138	0,033	0,172	0,001	0,069	0,079	-	1,000	0,220	0,160	1,000
Bangkok	1,000	1,000	0,914	1,000	1,000	1,000	1,000	-	0,866	1,000	0,591
Barcelona	0,358	0,253	0,048	0,219	0,289	0,299	0,220	0,866	-	0,380	1,000

Table 2. Distances between some cities according to the attribute Elevation

	Aberdeen	Abu_Dhabi	Agra	Amsterdam	Antwerp	Atlanta	Bahrain	Bangkok	Barcelona	Bath,_Some	Beijing
Aberdeen	-	0,017	0,599	0,009	0,008	1,000	0,021	0,006	0,078	0,047	0,154
Abu_Dhabi	0,017	-	0,616	0,026	0,025	1,000	0,004	0,023	0,096	0,064	0,171
Agra	0,599	0,616	-	0,591	0,592	0,599	0,621	0,594	0,521	0,552	0,446
Amsterdam	0,009	0,026	0,591	-	0,001	1,000	0,030	0,003	0,070	0,039	0,145
Antwerp	0,008	0,025	0,592	0,001	-	1,000	0,029	0,002	0,071	0,039	0,146
Atlanta	1,000	1,000	0,599	1,000	1,000	-	1,000	1,000	1,000	1,000	1,000
Bahrain	0,021	0,004	0,621	0,030	0,029	1,000	-	0,027	0,100	0,068	0,175
Bangkok	0,006	0,023	0,594	0,003	0,002	1,000	0,027	-	0,073	0,041	0,148
Barcelona	0,078	0,096	0,521	0,070	0,071	1,000	0,100	0,073	-	0,031	0,075

2.2 Categorical attributes

In the CITIES data matrix, the categorical attributes are two:

- Continent code – Categorical
- Climate – Categorical

In Deliverable D4 the Chi-squared distance is proposed for categorical attributes. This approach considers the frequencies of each category when calculating the distance to another category. Consequently, the underlying distribution of the modalities of the attribute influence the similarity values obtained. In particular, we have used a decomposition of the χ^2 metrics calculation prosed in (Gibert & Nonell, 2003).

$$d_k^2(i,i') = \begin{cases} 0, & if\ x_{ik} = x_{i'k} \\ \frac{1}{1_{k^i}} + \frac{1}{1_{k^{i'}}}, & otherwise \end{cases} \quad (5)$$

Table 3 shows the frequency distribution of the modalities for the two categorical attributes Continent and Climate.

Table 3. Frequency distribution of the Continent and Climate attributes.

Continent	Frequency	%
AF	3	2
AS	37	24,7
EU	79	52,7
NA	23	15,3
OC	2	1,3
SA	6	4

Climate	Frequency	%
Desert	6	4
Humid continental	15	10
Humid sub-tropical	33	22
Mediterranean	19	12,7
Oceanic	60	40
Semi-arid	2	1,3
Subarctic	1	0,7
Tropical monsoon	2	1,3
Tropical rainforest	2	1,3
Tropical savannah	10	6,7

Notice that if we calculate the distance for a city placed in EU and a city placed in SA we obtain:

$$d(cA, cB) = \frac{1}{79} + \frac{1}{6} = 0{,}18$$

And a larger distance is obtained when comparing a city placed in AF and another placed in OC:

$$d(cA, cB) = \frac{1}{3} + \frac{1}{2} = 0{,}83$$

These tables show some examples of the results obtained with the Chi-squared distance:

Table 4. Chi-squared distances between some cities according to the attribute Continent

	Aberdeen	Abu_Dhab	Agra	Amsterda	Antwerp	Atlanta	Bahrain	Bangkok	Barcelona	Bath,_Son	Beijing	Benidorm	Berlin	Bilbao
Aberdeen	-	0,040	0,040	0,000	0,000	0,056	0,040	0,040	0,000	0,000	0,040	0,000	0,000	0,000
Abu_Dhabi	0,040	-	0,000	0,040	0,040	0,071	0,000	0,000	0,040	0,040	0,000	0,040	0,040	0,040
Agra	0,040	0,000	-	0,040	0,040	0,071	0,000	0,000	0,040	0,040	0,000	0,040	0,040	0,040
Amsterdam	0,000	0,040	0,040	-	0,000	0,056	0,040	0,040	0,000	0,000	0,040	0,000	0,000	0,000
Antwerp	0,000	0,040	0,040	0,000	-	0,056	0,040	0,040	0,000	0,000	0,040	0,000	0,000	0,000
Atlanta	0,056	0,071	0,071	0,056	0,056	-	0,071	0,071	0,056	0,056	0,071	0,056	0,056	0,056
Bahrain	0,040	0,000	0,000	0,040	0,040	0,071	-	0,000	0,040	0,040	0,000	0,040	0,040	0,040
Bangkok	0,040	0,000	0,000	0,040	0,040	0,071	0,000	-	0,040	0,040	0,000	0,040	0,040	0,040
Barcelona	0,000	0,040	0,040	0,000	0,000	0,056	0,040	0,040	-	0,000	0,040	0,000	0,000	0,000

Table 5. Chi-squared distances between some cities according to the attribute Climate

	Aberdeen	Abu_Dhab	Agra	Amsterda	Antwerp	Atlanta	Bahrain	Bangkok	Barcelona	Bath,_Son	Beijing	Benidorm	Berlin	Bilbao
Aberdeen	-	0,183	0,517	0,000	0,000	0,047	0,183	0,117	0,069	0,000	0,083	0,069	0,000	0,000
Abu_Dhabi	0,183	-	0,667	0,183	0,183	0,197	0,000	0,267	0,219	0,183	0,233	0,219	0,183	0,183
Agra	0,517	0,667	-	0,517	0,517	0,530	0,667	0,600	0,553	0,517	0,567	0,553	0,517	0,517
Amsterdam	0,000	0,183	0,517	-	0,000	0,047	0,183	0,117	0,069	0,000	0,083	0,069	0,000	0,000
Antwerp	0,000	0,183	0,517	0,000	-	0,047	0,183	0,117	0,069	0,000	0,083	0,069	0,000	0,000
Atlanta	0,047	0,197	0,530	0,047	0,047	-	0,197	0,130	0,083	0,047	0,097	0,083	0,047	0,047
Bahrain	0,183	0,000	0,667	0,183	0,183	0,197	-	0,267	0,219	0,183	0,233	0,219	0,183	0,183
Bangkok	0,117	0,267	0,600	0,117	0,117	0,130	0,267	-	0,153	0,117	0,167	0,153	0,117	0,117
Barcelona	0,069	0,219	0,553	0,069	0,069	0,083	0,219	0,153	-	0,069	0,119	0,000	0,069	0,069

We can observe some "attraction" behaviour of the continents with high frequency. The goal of DAMASK recommender system is to help to diversify the destinations that are proposed to a tourist, so it seems not adequate that the cities that are in modalities with high concentration of options increase the similarity among them. For this reason, we finally have taken the Hamming distance based on the equality/inequality of the modalities. This distance takes into account the number of between the differences in the values of the categorical attributes, giving one of the following values when comparing the two modalities of the objects *i* and *i'* for the *k*-th attribute:

$$d'_k(i,i') = \begin{cases} 0 & if\ x_{ik} = x_{i'k} \\ 1 & if\ x_{ik} \neq x_{i'k} \end{cases} \quad (6)$$

Therefore, to calculate the overall distance for the set of categorical variables C, we make a weighted average of the partial distances given by Eq. 6 for each individual attribute, with $\sum_{k=1}^{|C!} w_k = 1$, as defined in Eq. 7.

$$d_k^C(i,i') = \sum_{k=1}^{|C|} w_k d'_k \quad (7)$$

The following tables show some examples of the results obtained with the Hamming distance.

Table 6. Hamming distance between some cities according to the attribute Continent

	Aberdeen	Abu_Dhab	Agra	Amsterda	Antwerp	Atlanta	Bahrain	Bangkok	Barcelona	Bath,_Son	Beijing	Benidorm	Berlin	Bilbao
Aberdeen	-	1,00	1,00	0,00	0,00	1,00	1,00	1,00	0,00	0,00	1,00	0,00	0,00	0,00
Abu_Dhabi	1,00	-	0,00	1,00	1,00	1,00	0,00	0,00	1,00	1,00	0,00	1,00	1,00	1,00
Agra	1,00	0,00	-	1,00	1,00	1,00	0,00	0,00	1,00	1,00	0,00	1,00	1,00	1,00
Amsterdam	0,00	1,00	1,00	-	0,00	1,00	1,00	1,00	0,00	0,00	1,00	0,00	0,00	0,00
Antwerp	0,00	1,00	1,00	0,00	-	1,00	1,00	1,00	0,00	0,00	1,00	0,00	0,00	0,00
Atlanta	1,00	1,00	1,00	1,00	1,00	-	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Bahrain	1,00	0,00	0,00	1,00	1,00	1,00	-	0,00	1,00	1,00	0,00	1,00	1,00	1,00
Bangkok	1,00	0,00	0,00	1,00	1,00	1,00	0,00	-	1,00	1,00	0,00	1,00	1,00	1,00
Barcelona	0,00	1,00	1,00	0,00	0,00	1,00	1,00	1,00	-	0,00	1,00	0,00	0,00	0,00

Table 7. Hamming distance between some cities according to the attribute Climate

	Aberdeen	Abu_Dhab	Agra	Amsterda	Antwerp	Atlanta	Bahrain	Bangkok	Barcelona	Bath,_Son	Beijing	Benidorm	Berlin	Bilbao
Aberdeen	-	1,00	1,00	0,00	0,00	1,00	1,00	1,00	1,00	0,00	1,00	1,00	0,00	0,00
Abu_Dhabi	1,00	-	1,00	1,00	1,00	1,00	0,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Agra	1,00	1,00	-	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Amsterdam	0,00	1,00	1,00	-	0,00	1,00	1,00	1,00	1,00	0,00	1,00	1,00	0,00	0,00
Antwerp	0,00	1,00	1,00	0,00	-	1,00	1,00	1,00	1,00	0,00	1,00	1,00	0,00	0,00
Atlanta	1,00	1,00	1,00	1,00	1,00	-	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Bahrain	1,00	0,00	1,00	1,00	1,00	1,00	-	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Bangkok	1,00	1,00	1,00	1,00	1,00	1,00	1,00	-	1,00	1,00	1,00	1,00	1,00	1,00
Barcelona	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	-	1,00	1,00	0,00	1,00	1,00

2.3 Semantic attributes

Finally, the third component of the compatibility distance measure corresponds to the contribution of the semantic features, which can be computed using any of the exiting measures presented in Deliverables D3 and D4. Those measures are based on the knowledge provided by the domain ontologies. In this case we will use the DAMASK ontology, explained in the Internal project report T3.2.

As argued in D4, the Superconcept-based distance (Batet, Valls, & Gibert, 2010) has been selected after the analysis of its behaviour in different datasets. The SCD definition for comparing a pair of concepts c_i and c_j is based on the following premises (see D4 for details):

- Let us define the full concept hierarchy or taxonomy (H^C) of concepts (C) of an ontology as a transitive is-a relation $H^C \in C \times C$.
- Let us define the set $\mathcal{A}(c_i)$ that contains the concept c_i and all the superconcepts (*i.e.*, ancestors) of c_i in a given taxonomy as:

$$\mathcal{A}(c_i)=\{c_j \in C \mid c_j \text{ is superconcept of } c_i\} \cup \{c_i\} \quad (8)$$

Then the Euclidean-based SuperConcept-based distance (SCD) is defined as

$$SCD(c_i, c_j) = \sqrt{\frac{|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)| - |\mathcal{A}(c_i) \cap \mathcal{A}(c_j)|}{|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)|}} \quad (9)$$

This is the squared root of the number of different ancestors divided by the number of total ancestors (the union). Let's calculate for instance the distance between *Church* and *Mosque* in the ontology represented in Figure 1.

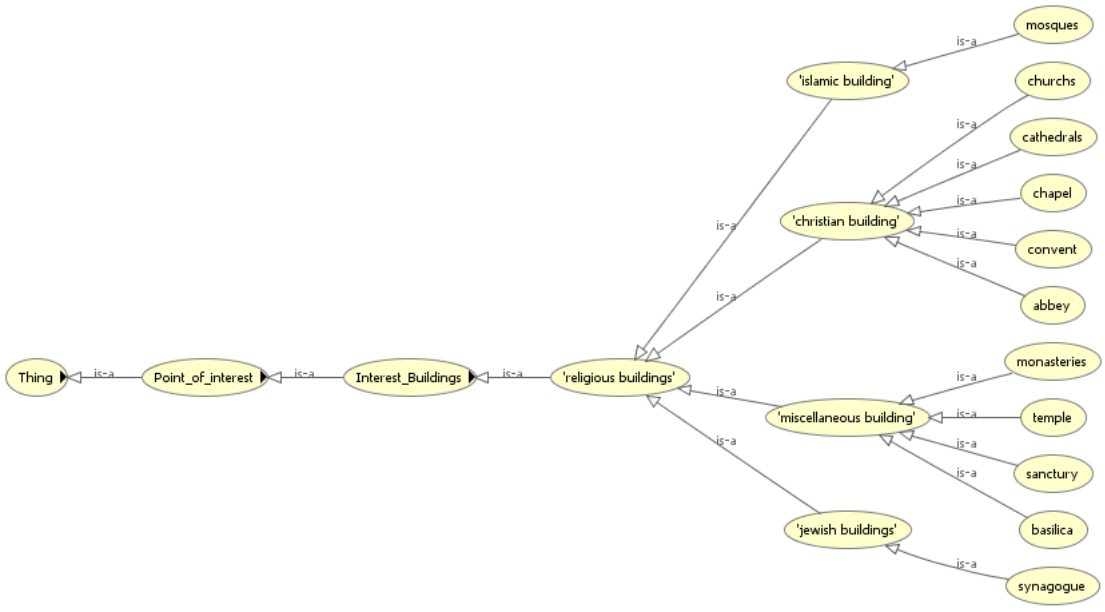


Figure 1: DAMASK ontology portion related to religious buildings

Church and mosque have 4 different ancestors (SCD requires counting itself as ancestor) and the union of ancestors is 8. Hence, the distance between Church and mosque is $\sqrt{4/8} = 0{,}7$. The squared root is used to smooth the result and highlight the small differences.

2.3.1 An extension of SCD to be used in multi-valued attributes

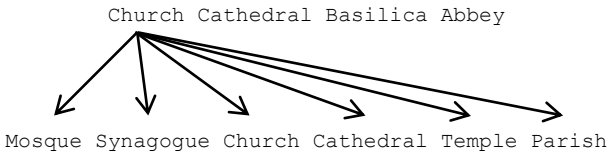
In the CITIES data matrix the attributes corresponding to semantic features usually have more than one value (see the Internal Report T3-2). Therefore, the case of multi-valued semantic attributes has been studied. Let us consider an example: calculate the distance between Barcelona and Berlin regarding their religious buildings. The values of these cities are the following. According to the elicitation process we know that all the values correspond to some concept in the DAMASK ontology (so we do not require considering the case of values that are not found in the ontology).

Barcelona	`Church#Cathedral#Basilica#Abbey`
Berlin	`Mosque#Synagogue#Church#Cathedral#Temple#Parish`

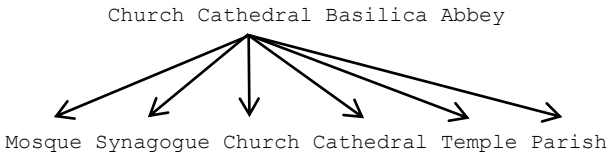
The algorithm used to determine the distance between two cities is as follows:

1. Take a concept value of city A and calculate its semantic distance to each concept of city B, using the ontology. This results in an array of distances.
2. Take the minimum distance on this array and register it in an auxiliary array. In the example above, the minimum distance of Church (in Barcelona) to all the values in Berlin is 0.0 (Church also in Berlin).
3. Repeat this process for all the concepts in A with respect to B.
4. Repeat this process for all the concepts in B with respect to A.
5. Aggregate all the partial distance values with the Ordered Weighted Average (OWA) operator.

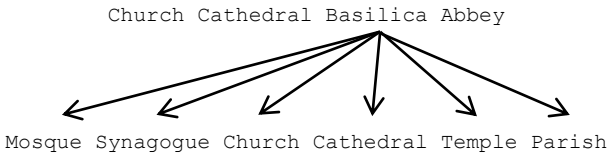
Let us consider the example of comparing Barcelona and Berlin. The first step makes the following calculations:



This will result in an array of distances like these: 0.7, 0.7, 0.0, 0.2, 0.7 and 0.9. The minimum is 0.0.

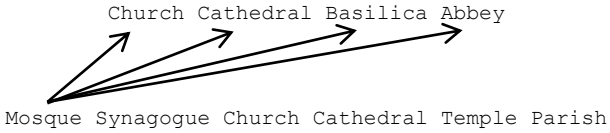


As before, both cities have a cathedral, so the distance here to save is also 0.0. Now in the array we have [0.0, 0.0].



Now, this results in an array of distances like these: 0.7, 0.7, 0.7, 0.7, 0.2 and 0.9. The minimum distance now is between Basilica and Temple (0.2); we save it in the array -> [0.0, 0.0, 0.2].

The process continues for all the concepts for city A, and then we do the same from city B to city A:



When all the distances are calculated, we will end with an array like this: [0.0, 0.0, 0.2, 0.2, 0.7, 0.7, 0.0, 0.0, 0.2, 0.9]. Next is to apply the operator OWA to this array.

This process can be summarised in the following definition.

Definition 1: SuperConcept-based distance for multi-valued attributes (SCD_{mv})

$$SCD_{mv}(i,i') = OWA_{\omega}\left(\left\{\forall c_i : min_{\forall c_{i'}}\left(SCD(c_i, c_{i'})\right)\right\} \cup \left\{\forall c_{i'} : min_{\forall c_i}\left(SCD(c_i, c_{i'})\right)\right\}\right) \quad (10)$$

Finally, the distance for semantic attributes that is used in the compatibility measure is a weighted average of the SuperConcept distances obtained for each of the attributes:

$$d_k^S(i,i') = \sum_{k=1}^{|S|} w_k \, SCD_{mv}(i,i') \quad (11)$$

This approach to multi-valued data is based on the aggregation operation OWA. This operator was defined by R.R. Yager in (Yager, 1988). Since its appearance, it has been studied by many authors and it has been widely applied to many decision making problems (Beliakov, Pradera, & Calvo, 2007; Herrera, Herrera-Viedma, & Verdegay, 1996; Merigo & Gil-Lafuente, 2009; Xu, 2006).

Definition 2: A function $F: R^n \rightarrow R$ is an *OWA* operator of dimension n if it has an associated vector ω of dimension n such that its components satisfy:

a. $\omega_j \in [0,1]$
b. $\sum_{j=1}^{n} \omega_j = 1$

And:

$$F(a_1, a_2, \ldots, a_n) = \sum_{j=1}^{n} \omega_j b_j \quad (12)$$

, where b_j is the *j-th* largest element of the bag $\langle a_1, a_2, \ldots, a_n \rangle$.

Notice that the fundamental aspect of this operator is the re-ordering step, in particular an argument a_i is not associated with a particular weight w_i but rather a weight is associated with a particular ordered position of argument.

The set of weights is extremely important in the OWA method, because it determines the aggregation policy that the decision maker is imposing on the decision process. Some measures have been introduced to characterize a weight vector, such as evaluating its **attitudinal-character** (or **orness**), which is defined as (Yager, 1988):

$$\alpha(\omega) = \frac{1}{n-1} \sum_{i=0}^{n} \omega_i \, (n - i) \quad (13)$$

It is known that $\alpha \in [0,1]$. As a general rule, as the allocation of weight in W moves to the top, then α gets closer to one, meanwhile as the weights move to the bottom, α gets closer to zero. Furthermore, if W is symmetrical, then $\alpha(\omega) = 0.5$. This measure provides a characterization of the type of aggregation being performed. An α value near one indicates a bias toward considering mainly the larger values in the argument (i.e. high **orness** or disjunctive behaviour), while an α value near zero indicates preference is being given to the smaller values in the argument (i.e. high **andness** or conjunctive behaviour). An α value near 0.5 is an indication of a neutral type aggregation (i.e. averaging).

2.3.2 Generation of the OWA weights

When comparing pairs of cities, the number of arguments (i.e. partial distances) to aggregate is not a constant, it depends on the number of concepts that each city has, consequently we cannot use predetermined OWA weights. For example, if one city has 2 religious buildings and it is compared with another city with 5 religions buildings, with the process described in xx, we will generate 2 × 5 = 10 similarity values to be aggregated using the OWA operator.

In this section, we analyse three different methods for generating automatically the set of OWA weights.

1. **Borda-Kendall law:** which uses a linear function (Lamata & Cables, 2009; Lamata & Pérez, 2012):

$$\omega_i = \frac{2(n+1-i)}{n(n+1)} \quad (14)$$

The resulting array of weights for *n=10* is: [0.182, 0.164, 0.145, 0.127, 0.109, 0.091, 0.073, 0.055, 0.036, 0.018]. This is the graphic of weights:

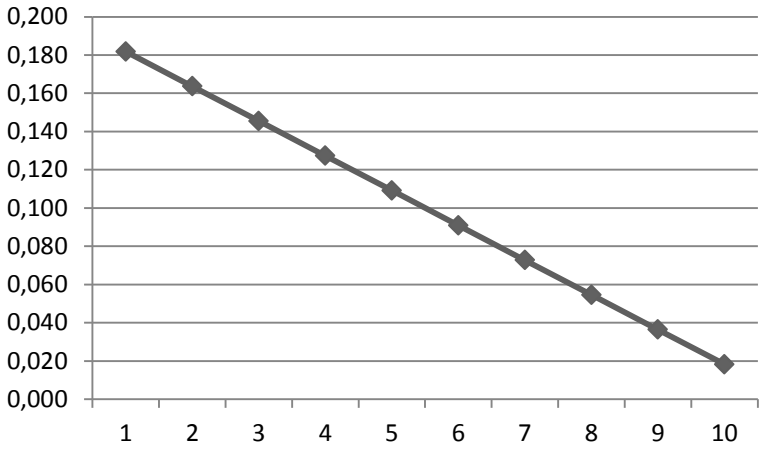


Figure 2: Graphic of the resulting OWA weights using a linear function

If we apply these weights to the ordered (from lower to higher) array of distances and sum them all, we obtain the distance between the two cities. Following the previous example, for the *religious buildings* attribute, the distance between Barcelona and Berlin: 0.134

Notice that this method gives more weight to the similarities than to the differences. It has high *orness* with a value of 0.67.

This process is executed by a Java application that will result in various excel files (Table 8), one per each column in the main data matrix, containing all the distances between each of the cities.

Table 8. Example of the result for the "Religious buildings" attribute

	Aberdeen	Abu_Dhab	Agra	Amsterdar	Antwerp	Atlanta	Bahrain	Bangkok	Barcelona	Bath._Son	Beijing	Benidorm	Berlin	Bilbao
Aberdeen	-	0,18	0,35	0,48	0,16	0,05	0,18	0,98	0,21	0,15	0,23	0,75	0,05	0,33
Abu_Dhab	0,18	-	0,21	0,39	0,44	0,05	0,00	0,39	0,45	0,31	0,10	0,75	0,05	0,42
Agra	0,35	0,21	-	0,83	0,44	0,15	0,21	0,34	0,43	0,30	0,10	0,75	0,15	0,40
Amsterdar	0,48	0,39	0,83	-	0,22	0,49	0,39	0,97	0,31	0,45	0,42	0,75	0,49	0,24
Antwerp	0,16	0,44	0,44	0,22	-	0,30	0,44	0,97	0,03	0,14	0,24	0,75	0,30	0,12
Atlanta	0,05	0,05	0,15	0,49	0,30	-	0,05	0,53	0,31	0,10	0,08	0,75	0,00	0,28
Bahrain	0,18	0,00	0,21	0,39	0,44	0,05	-	0,39	0,45	0,31	0,10	0,75	0,05	0,42
Bangkok	0,98	0,39	0,34	0,97	0,97	0,53	0,39	-	0,83	0,49	0,42	0,75	0,53	0,80
Barcelona	0,21	0,45	0,43	0,31	0,03	0,31	0,45	0,83	-	0,14	0,25	0,75	0,31	0,03
Bath._Son	0,15	0,31	0,30	0,45	0,14	0,10	0,31	0,49	0,14	-	0,08	0,75	0,10	0,26
Beijing	0,23	0,10	0,10	0,42	0,24	0,08	0,10	0,42	0,25	0,08	-	0,75	0,08	0,21
Benidorm	0,75	0,75	0,75	0,75	0,75	0,75	0,75	0,75	0,75	0,75	0,75	-	0,75	0,75
Berlin	0,05	0,05	0,15	0,49	0,30	0,00	0,05	0,53	0,31	0,10	0,08	0,75	-	0,28
Bilbao	0,33	0,42	0,40	0,24	0,12	0,28	0,42	0,80	0,03	0,26	0,21	0,75	0,28	-
Birminghar	0,05	0,05	0,15	0,49	0,30	0,00	0,05	0,53	0,31	0,10	0,08	0,75	0,00	0,28
Boston	0,48	0,39	0,83	0,00	0,22	0,49	0,39	0,97	0,31	0,45	0,42	0,75	0,49	0,24
Bratislava	0,33	0,42	0,40	0,24	0,12	0,28	0,42	0,80	0,16	0,11	0,07	0,75	0,28	0,10
Bregenz	0,11	0,49	0,75	0,40	0,23	0,32	0,49	0,86	0,26	0,07	0,30	0,75	0,32	0,44

2. **Non-linear decreasing function**. We have defined a function that generates a set of weights in a non-linear decreasing way as shown in Figure 3. The equation used is the following:

$$\omega_i = \frac{1}{x^{\frac{4}{5}}} \tag{15}$$

This will result in an array of weights that do not sum 1. To solve this, all the resulting values are summed and then each one is divided by this resulting sum, like normalization. Now the weights sum 1 as expected, and this are their values for *n=10*: [0.281 0.161 0.116 0.093 0.077 0.067 0.059 0.053 0.048 0.044]. Its *orness* is 0.69 and this is the graphic of weights:

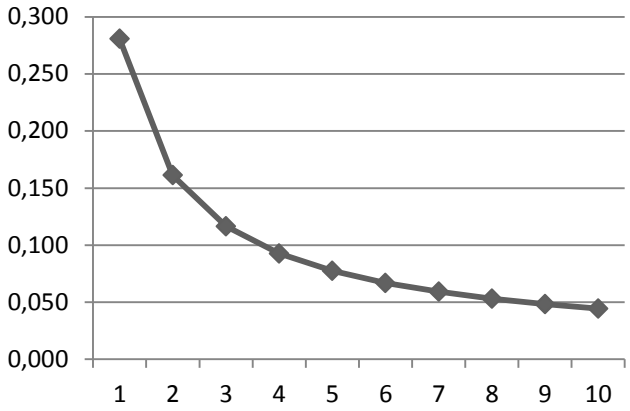


Figure 3: Graphic of the resulting weights using a non-linear decreasing function

Table 9. Example of the result for the "Religious buildings" attribute using the non-linear decreasing set of weights for OWA

	Aberdeen	Abu_Dhab	Agra	Amsterdar	Antwerp	Atlanta	Bahrain	Bangkok	Barcelona	Bath._Son	Beijing	Benidorm	Berlin	Bilbao
Aberdeen	-	0,20	0,33	0,44	0,20	0,08	0,20	0,97	0,23	0,17	0,24	0,75	0,08	0,31
Abu_Dhab	0,20	-	0,22	0,38	0,40	0,09	0,00	0,38	0,40	0,30	0,13	0,25	0,09	0,38
Agra	0,33	0,22	-	0,81	0,40	0,18	0,22	0,34	0,38	0,29	0,12	0,25	0,18	0,36
Amsterdar	0,44	0,38	0,81	-	0,23	0,45	0,38	0,95	0,31	0,43	0,39	0,75	0,45	0,26
Antwerp	0,20	0,40	0,40	0,23	-	0,30	0,40	0,95	0,07	0,18	0,25	0,25	0,30	0,15
Atlanta	0,08	0,09	0,18	0,45	0,30	-	0,09	0,48	0,30	0,13	0,12	0,25	0,00	0,29
Bahrain	0,20	0,00	0,22	0,38	0,40	0,09	-	0,38	0,40	0,30	0,13	0,25	0,09	0,38
Bangkok	0,97	0,38	0,34	0,95	0,95	0,48	0,38	-	0,81	0,45	0,39	0,75	0,48	0,80
Barcelona	0,23	0,40	0,38	0,31	0,07	0,30	0,40	0,81	-	0,17	0,24	0,25	0,30	0,05
Bath._Son	0,17	0,30	0,29	0,43	0,18	0,13	0,30	0,45	0,17	-	0,12	0,25	0,13	0,26
Beijing	0,24	0,13	0,12	0,39	0,25	0,12	0,13	0,39	0,24	0,12	-	0,75	0,12	0,22
Benidorm	0,75	0,25	0,25	0,75	0,25	0,25	0,25	0,75	0,25	0,25	0,75	-	0,25	0,25
Berlin	0,08	0,09	0,18	0,45	0,30	0,00	0,09	0,48	0,30	0,13	0,12	0,25	-	0,29
Bilbao	0,31	0,38	0,36	0,26	0,15	0,29	0,38	0,80	0,05	0,26	0,22	0,25	0,29	-
Birminghar	0,08	0,09	0,18	0,45	0,30	0,00	0,09	0,48	0,30	0,13	0,12	0,25	0,00	0,29
Boston	0,44	0,38	0,81	0,00	0,23	0,45	0,38	0,95	0,31	0,43	0,39	0,75	0,45	0,26
Bratislava	0,31	0,38	0,36	0,26	0,15	0,29	0,38	0,80	0,17	0,15	0,10	0,25	0,29	0,13
Bregenz	0,14	0,44	0,76	0,39	0,25	0,29	0,44	0,85	0,26	0,09	0,29	0,75	0,29	0,40

These results are similar to the ones seen with the previous method, which only small changes. This is due to the bigger weight given to the first value and to the major dissimilarities.

3. **Linguistic quantifiers:** The classical logic uses just two quantifiers, which are: the universal quantifier ∀ (all) and the existential quantifier ∃ (exists). But one may want to use something in between like *most*, *many*, *at least half*, *some*, and *few*. These are the linguistic quantifiers defined for fuzzy sets in (Yager, 1993, 1996). They permit to model different compensation aggregation mechanism, applied to define different behavioural policies (from pessimistic – conjunctive – to optimistic – disjunctive).

   After a careful study of the behavioural character of the different approaches, we have decided to use the linguistic quantifier *many* for aggregating the partial distances obtained for semantic attributes. This quantifier models a partial conjunctive policy. It means that a city will be similar to another one if *many* of their values in the semantic attributes are similar. Not permitting the compensation of low values with high ones.

   The weights associated to linguistic quantifiers are usually obtained from Fuzzy Quantifiers (Yager, 1996). A function $Q : [0, 1] \rightarrow [0, 1]$ *is a* regular monotonically non-decreasing fuzzy quantifier *(non-decreasing fuzzy quantifiers for short)* if it satisfies:

(i) $Q(0) = 0$;
(ii) $Q(1) = 1$;
(iii) $x > y$ implies $Q(x) \geq Q(y)$.

A well-known fuzzy quantifier is based on the sigmoidal function and it is given by the following definition.

$$Q^{\alpha}(x) = f(x) = \begin{cases} 0, & if\ x = 0 \\ \frac{1}{1 + e^{(\alpha - x)*10}} \text{ for } \alpha > 0, & if\ 0 < x < 1 \\ 1, & if\ x = 1 \end{cases} \quad (15)$$

A graphical representation of this fuzzy quantifier is given in Figure 4 for some particular values on the parameter α, α= {0, 0.1, . . . 0.9}. We can observe that for small a values, the function increases quickly near $x = 0$,whereas the increase is smoothly for larger values of α.

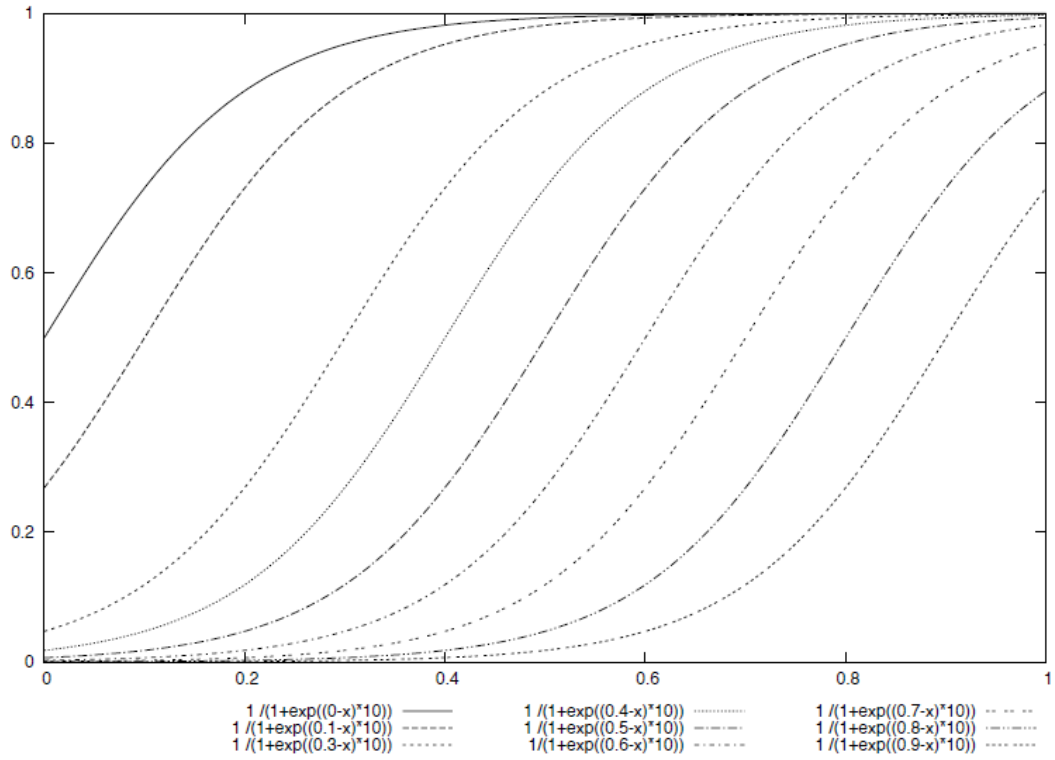


Figure 4: Representation of function 15 for α 0 to 0.9.

Using this fuzzy quantifier, the OWA weights can be obtained with the following equation:

$$w_i = \left[Q\left(\frac{i}{N}\right) - Q\left(\frac{i-1}{N}\right) \right] \quad (16)$$

For the linguistic quantifier "most", the recommended value is α = 0.6. Using Eq 16 and Eq 15, the set of weights obtained to aggregate 10 values is shown in Figure 5. Taking into account that the OWA operator will sort the values in a decreasing way, we are giving high weights for those values that are below the median, which assures that all the previous ones are equal or higher. Different values of the parameter α would shift the curve in the figure 5 to the left for lower values and to the right otherwise.

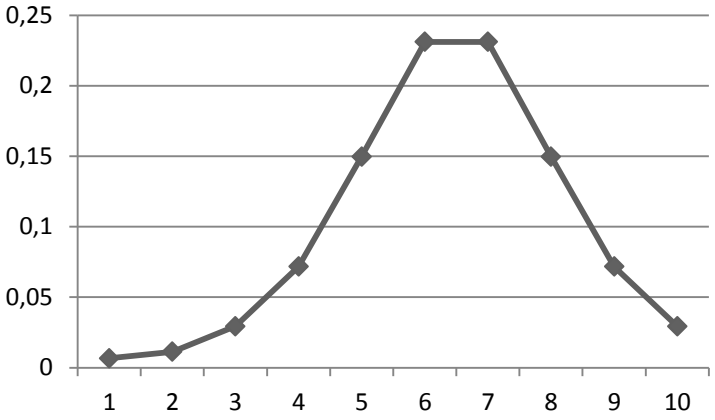


Figure 5: Graphic of the resulting weights using the linguistic quantifier defined by Eq. 16 with α = 0.6 for n = 10

The results obtained with this set of weights in the OWA operator are quite different from the ones resulting when applying the previous methods: for *n=10* the resulting weights are [0.007 0.011 0.029 0.072 0.150 0.231 0.231 0.150 0.072 0.029] and its *orness* is 0.39. In this case, the weight for the most similar and the most dissimilar concepts of the array are low, so that if most of the cities do only coincide in 1 value (f.i. in religious buildings, almost every city has a church), the result is a high value of similarity using the methods 1 and 2, but not with this one.

Here is another example represented in Figure 6 for *n=4*, which results in weights [0.029 0.239 0.548 0.164] and an *orness* of 0.28.

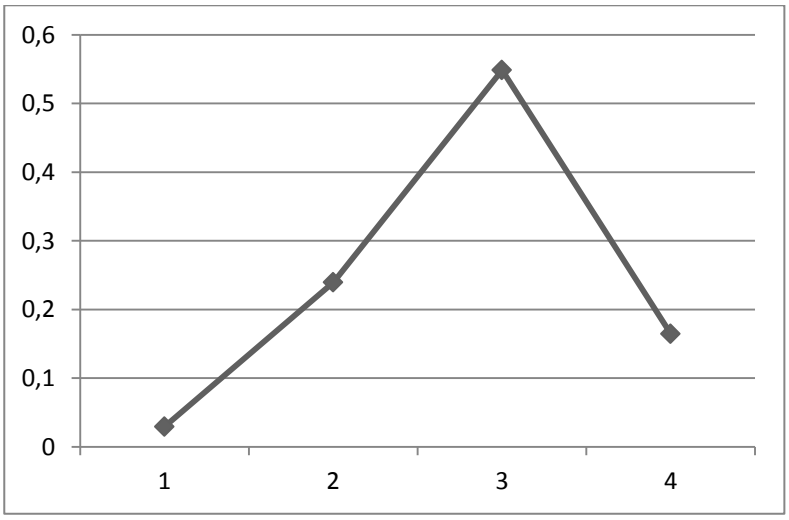


Figure 6: Graphic of the resulting weights using the linguistic quantifier defined by Eq. 16 with α = 0.6 for n = 4

It can be seen that as explained before, the most similar and the most dissimilar values have low weights. The orness values obtained indicate that in this case we are taking a more conjunctive behaviour, less compensative. With this approach we are able to stress the differences, enhancing the discriminating power of semantic features. This is an interesting result for clustering purposes, so this third approach is the one that will be included in the recommender system.

Table 10. Example of the results for the attribute "Religious buildings" using the linguistic quantifier "most".

	Aberdeen	Abu_Dhab	Agra	Amsterdar	Antwerp	Atlanta	Bahrain	Bangkok	Barcelona	Bath._Son	Beijing	Benidorm	Berlin	Bilbao
Aberdeen	-	0,48	0,75	0,81	0,42	0,10	0,48	0,96	0,55	0,39	0,60	0,75	0,10	0,75
Abu_Dhab	0,48	-	0,53	0,83	0,82	0,09	0,00	0,83	0,73	0,70	0,24	0,75	0,09	0,74
Agra	0,75	0,53	-	0,92	0,82	0,38	0,53	0,74	0,70	0,66	0,23	0,75	0,38	0,70
Amsterdar	0,81	0,83	0,92	-	0,52	0,90	0,83	0,95	0,65	0,83	0,84	0,75	0,90	0,56
Antwerp	0,42	0,82	0,82	0,52	-	0,74	0,82	0,95	0,05	0,32	0,62	0,75	0,74	0,25
Atlanta	0,10	0,09	0,38	0,90	0,74	-	0,09	0,92	0,70	0,25	0,17	0,75	0,00	0,68
Bahrain	0,48	0,00	0,53	0,83	0,82	0,09	-	0,83	0,73	0,70	0,24	0,75	0,09	0,74
Bangkok	0,96	0,83	0,74	0,95	0,95	0,92	0,83	-	0,92	0,90	0,84	0,75	0,92	0,88
Barcelona	0,55	0,73	0,70	0,65	0,05	0,70	0,73	0,92	-	0,37	0,60	0,75	0,70	0,04
Bath._Son	0,39	0,70	0,66	0,83	0,32	0,25	0,70	0,90	0,37	-	0,17	0,75	0,25	0,61
Beijing	0,60	0,24	0,23	0,84	0,62	0,17	0,24	0,84	0,60	0,17	-	0,75	0,17	0,53
Benidorm	0,75	0,75	0,75	0,75	0,75	0,75	0,75	0,75	0,75	0,75	0,75	-	0,75	0,75
Berlin	0,10	0,09	0,38	0,90	0,74	0,00	0,09	0,92	0,70	0,25	0,17	0,75	-	0,68
Bilbao	0,75	0,74	0,70	0,56	0,25	0,68	0,74	0,88	0,04	0,61	0,53	0,75	0,68	-
Birminghar	0,10	0,09	0,38	0,90	0,74	0,00	0,09	0,92	0,70	0,25	0,17	0,75	0,00	0,68
Boston	0,81	0,83	0,92	0,00	0,52	0,90	0,83	0,95	0,65	0,83	0,84	0,75	0,90	0,56
Bratislava	0,75	0,74	0,70	0,56	0,25	0,68	0,74	0,88	0,40	0,25	0,13	0,75	0,68	0,23
Bregenz	0,27	0,79	0,76	0,77	0,57	0,67	0,79	0,95	0,61	0,15	0,66	0,75	0,67	0,72

2.3.2.1 Comparison of the different methods for generating OWA weights

In order to apreciate the differences between the three different techniques for generating automatically the set of weights, a comparative has been done in table 11.

Table 11. Comparative of the 3 methods for the "Religious buildings" attribute.

	Linear OWA	Non-linear OWA	Linguistic qualifier OWA
Aberdeen – Agra	0.35	0.33	0.75
Antwerp – Barcelona	0.03	0.07	0.05
Agra – Amsterdam	0.83	0.81	0.92
Aberdeen – Bangkok	0.98	0.97	0.96
Barcelona – Budapest	0.11	0.15	0.26

It is easy to see that the distances are a bit larger if the linguistic qualifier *most* is used, because it is more pessimistic than the two first approaches. This specially noted between Aberdeen and Agra, because we have decreased the compensation factor. For cities with a lot of values in common, such as Antwerp and Barcelona, the difference is small. Similarly, with cities with very few things in common, such as Aberdeen and Bangkok, the three approaches give also a quite similar distance value. Further analysis will be done using the results of the clustering.

2.4 Treatment of the missing values

The treatment of missing values must deserve special attention in clustering algorithms. In the CITIES data matrix that has been compiled in the DAMASK project we only find missing values in the semantic attributes. The numerical and categorical information is complete for all the cities, because we have used different extraction mechanisms until obtaining all the data, as explained in the Internal Report T3-2.

For the case of semantic descriptions, in case that the system is not able to find any evidence for a given attribute, the symbol '?' is used. As explained in the Internal Report T3-2, in this case, the symbol is not exactly representing a missing value as normally understood, because the lack of information about one attribute is telling us that probably the city does not have any instance of this type. For example, in Figure 7, the automatic extraction system has not found any information about Maritime Museums in Munich, because certainly they do not exist. The data matrix construction by means of extraction processes is slanted by the *precision* and *recall* of each method used during the whole process (i.e., the natural language parser, the named entity detection heuristics, the inaccuracy of Web statistics and the relat-edness measures). The *precision index* measures the number of correct values among all the values obtained. *Recall* is calculated by dividing the number of correct values by the total of values that could have been found. As explained in deliverable D2, for the purpose of the project, high precision is needed, to ensure that the values that we attach to some city are correct. High precision is achieven at a cost of reducing the recall. In this case, the symbol '?' may appear in the data matrix because we have not been able to retrieve the information from the Web page. For example Oslo has a Natural History Museum, but this data has not been found by the system.

Matrix

City-name	Water_Landmark	Geographical_La...	Natural_History_Mu...	Maritime_Muse...	Christian_Buildi...
Mumbai	?	?	?	?	?
Munich	River#Bridge	Square#Hill#Mou...	Natural_History_M...	?	Cathedral#Abb...
Nanjing	Lake#River#Bridge	Hill#Mountain	Natural_History_M...	?	Church
Naples	?	Cave#Square#Hi...	?	?	Chapel#Church...
New_Delhi	?	Hill#Mountain	?	?	Abbey
New_York_C...	Canal#Beach#Riv...	Square#Hill#Mou...	Natural_History_M...	?	Abbey
Newcastle_u...	Canal#River#Brid...	Gorge#Square#...	Natural_History_M...	?	Abbey
Nice	Beach#River#Brid...	Square#Hill#Terr...	Natural_History_M...	?	Church#Cathed...
Nottingham	Canal#Lake#River...	Cave#Square#Hill	Natural_History_M...	?	Chapel#Church...
Nuremberg	Canal#River#Bridge	?	?	?	Chapel#Church...
Orlando,_Flo...	Beach#Lake#Rive...	Square#Hill#Terr...	?	?	Chapel#Church
Oslo	Lake#River	Square#Hill#Mou...	?	Maritime_Muse...	Cathedral#Chur...
Oxford	Canal#River#Bridge	Square#Hill	Natural_History_M...	?	Chapel#Church...
Paris	Canal#River#Bridge	Square#Hill	?	?	Chapel#Church...
Prague	Bridge#Lake#River	Square#Hill#Mou...	Natural_History_M...	?	Chapel#Church...
Qingdao	Beach#River	Square#Hill#Mou...			Church#Cathed...
Reading,_Pe...	River#Canal	Square#Mountain	?	?	Church
Reading,_Be...	Canal#Lake#River...	Hill	?	?	Church#Abbey#...
Rheims	Canal	Cave#Square#Hill			Chapel#Church...
Reykjavík	Beach#Lake	Hill#Mountain	?	?	Church
Rio_de_Jan...	Beach#Lake#Rive...	Square#Hill#Mou...	Natural_History_M...	?	Chapel
Rome	River#Bridge	Square#Hill			Chapel#Church...
Saint_Peters...	Canal#Lake#River...	Square#Hill	Natural_History_M...	Maritime_Muse...	Church#Cathed...
Salvador,_Ba...	Beach#Lake#River	Square	?	?	Church#Cathed...
Salzburg	Lake#River#Bridge	Hill#Mountain			Cathedral#Abb...
San_Diego	Canal#Beach	Hill#Terrace#Mo...	Natural_History_M...	Maritime_Muse...	Church

Figure 7: Resulting matrix from the Tree extracting procedure

The distance for a city that has missing data value to another city with known data has been established to a value of 0.75. A value higher than 0.5 has been fixed in order to represent that a city with something is far from a city with probably no elements of the same typology. Due to the recall error, as a missing cannot absolutly mean that the city has no instances for the attribute, the value of distance used is not 1, but 0.75.

Moreover, the distance between two cities that have missings has been set to 0.25. In this case, this lower value on the distance (i.e. higher similarity) represents that these two cities have something in common as both may be cities without instances on the given attribute. Again, due to the recall error, the distance is set to 0.25 and not 0.

Table 11. Distances applied to semantic values when missing information

	'?'	**No missing**
'?'	0.25	0.75
No missing	0.75	Calculated with eq.10

3 References

Anderberg, M. R. (1973). *Cluster analysis for applications* (p. 359). Academic Press.

Batet, M. (2010). *Ontology-based semantic clustering*. URV. Retrieved from http://deim.urv.cat/~itaka/CMS/images/pdf/tesis_mbatetdeim.pdf

Batet, M., Valls, A., & Gibert, K. (2010). A distance function to assess the similarity of words using ontologies. *Proceeding of the XV congreso español sobre tecnologías y lógica fuzzy, Huelva* (pp. 561–566). Retrieved from http://www.uhu.es/estylf2010/trabajos/SS10-01.pdf

Batet, M., Valls, A., & Gibert, K. (2011). Semantic Clustering based on Ontologies - An Application to the Study of Visitors in a Natural Reserve. *ICAART (1)* (pp. 283-289). Retrieved from http://dblp.uni-trier.de/db/conf/icaart/icaart2011-1.html#BatetVG11

Beliakov, G., Pradera, A., & Calvo, T. (2007). *Aggregation Functions: A Guide for Practitioners* (p. 361).

Gibert, K., & Nonell, R. (2003). Impact of Mixed Metrics on Clustering. *Lecture Notes in Computer Science*, *2905/2003*, 464-471. doi:/10.1007/978-3-540-24586-5_57

Herrera, F., Herrera-Viedma, E., & Verdegay, J. L. (1996). Direct approach processes in group decision making using linguistic OWA operators. *Fuzzy Sets and Systems*, *79*(2), 175-190. doi:10.1016/0165-0114(95)00162-X

Lamata, M. T., & Cables, E. (2009). OWA weights determination by means of linear functions. *Mathware & Soft Computing*, *16*, 107-122. Retrieved from http://ic.ugr.es/mathware/index.php/Mathware/article/view/398

Lamata, M. T., & Pérez, E. C. (2012). Obtaining OWA operators starting from a linear order and preference quantifiers. *International Journal of Intelligent Systems*, *27*(3), 242-258. doi:10.1002/int.21520

Merigo, J., & Gil-Lafuente, A. (2009). The induced generalized OWA operator. *Information Sciences*, *179*(6), 729-741. Elsevier Inc. doi:10.1016/j.ins.2008.11.013

Xu, Z. (2006). Dependent OWA operators. *Lecture Notes in Computer Science*, *3885/2006*, 172-178. doi:10.1007/11681960_18

Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics*, *18*(1), 183-190. doi:10.1109/21.87068

Yager, R. R. (1993). Families of OWA operators. *Fuzzy Sets and Systems*, *59*(2), 125-148. Elsevier. doi:10.1016/0165-0114(93)90194-M

Yager, R. R. (1996). Quantifier guided aggregation using OWA operators. *International Journal of Intelligent Systems*, *11*(1), 49-73. doi:10.1002/(SICI)1098-111X(199601)11:1<49::AID-INT3>3.3.CO;2-L