Affinity of ODL_I^3 classes

Hierarchical clustering techniques based on the concept of *affinity*.

Affinity coefficients to determine the level of similarity between two classes in different source schemas.

An affinity function *A()* is defined on top of the Common Thesaurus to evaluate the affinity of two terms.

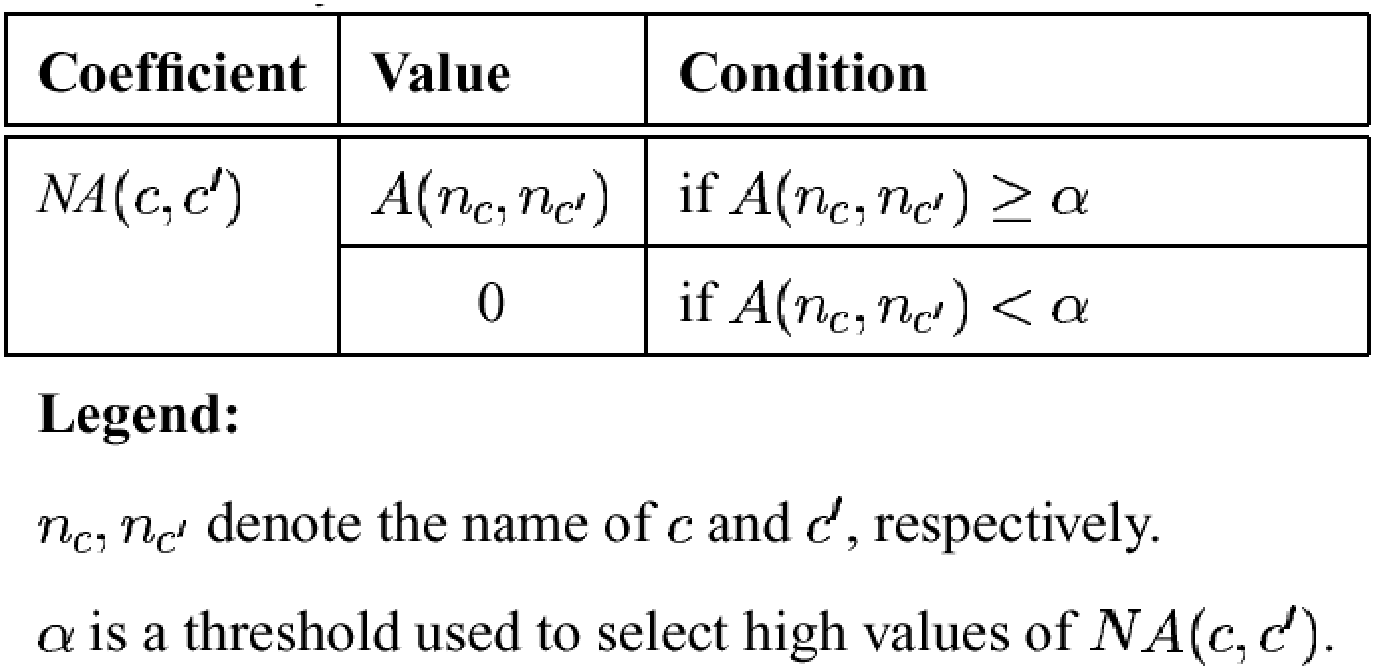
a strength σr is assigned to each type of relationship in the Common Thesaurus, with σr(SYN) ≥ σr(BT/NT) ≥ σr(RT)

The affinity *A(t,t')* of two terms *t* and *t'* is equal to the highest-strength path of relationships between them, if at least one path exist, and is zero otherwise.

~ denote that two terms have affinity.

Name Affinity coefficient - NA

Is the measure of the affinity between name of classes *c,c* ,
if this measure exceeds a specified threshold



Coefficient	**Value**	**Condition**
$NA(c, c')$	$A(n_c, n_{c'})$	if $A(n_c, n_{c'}) \geq \alpha$
	0	if $A(n_c, n_{c'}) < \alpha$

Legend:

$n_c, n_{c'}$ denote the name of c and c', respectively.

α is a threshold used to select high values of $NA(c, c')$.

Structural Affinity Coefficient - SA

Is the measure of the level of matching of two classes *c,c* based on attribute relationships in the Common Thesaurus:

$$SA(c,c') = \frac{2\cdot \mid \{(a_t, a_q) \mid a_t \in A(c), a_q \in A(c'), n_t \sim n_q\} \mid}{\mid A(c) \mid + \mid A(c') \mid} \cdot F_c$$

$$F_c \quad = \quad \frac{|\{x \in C| \ flag(x)=1 \ \}|}{|C|}$$

$$\mathrm{C} \quad = \quad \{(a_t, a_q) \mid a_t \in A(c), a_q \in A(c'), \langle a_t \text{ SYN } a_q\rangle \text{ or } \langle a_t \text{ BT } a_q\rangle \text{ or } \langle a_t \text{ NT } a_q\rangle\}$$

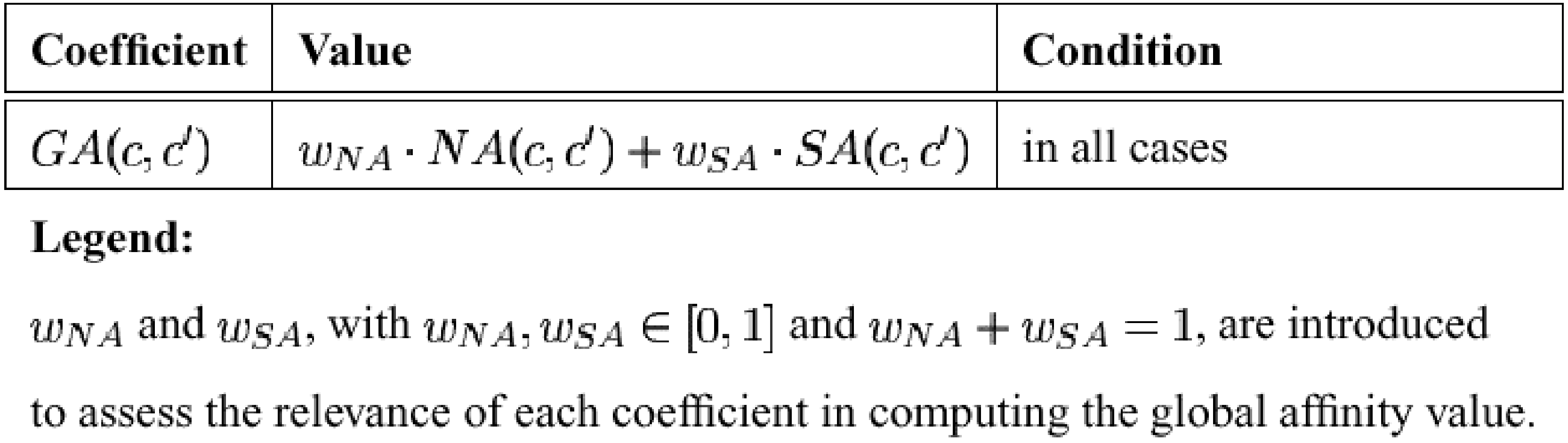
Legend

A(C) is the set of attributes in *C*

Flag(x)=1 stands for a valid relationship in the Common Thesaurus

Global Affinity Coefficient - GA

Is the measure of the affinity between classes *c,c*
computed as the weighted sum of the *NA* and *SA* coefficients



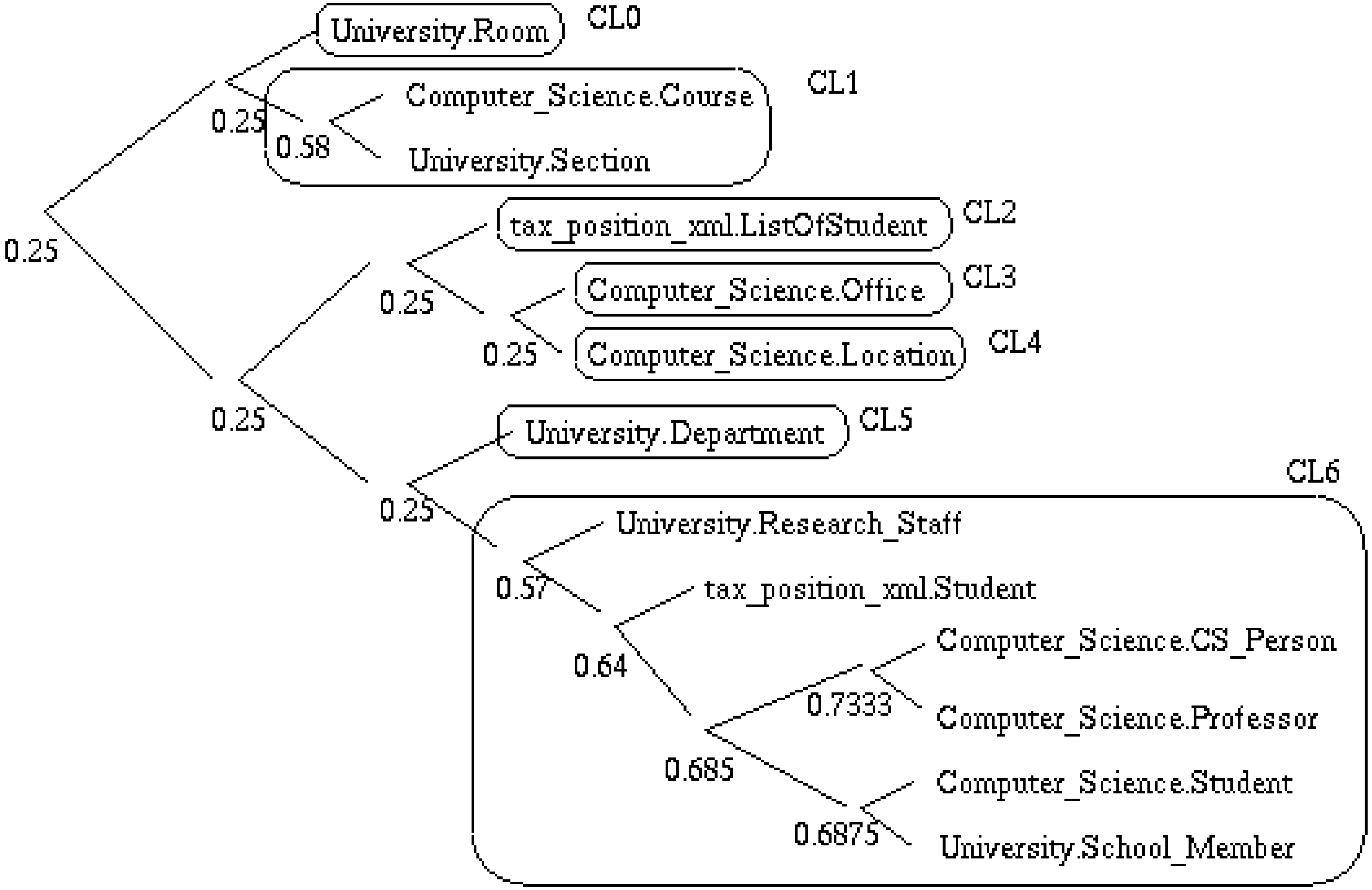
Coefficient	**Value**	**Condition**
$GA(c, c')$	$w_{NA} \cdot NA(c, c') + w_{SA} \cdot SA(c, c')$	in all cases

Legend:

w_{NA} and w_{SA}, with $w_{NA}, w_{SA} \in [0, 1]$ and $w_{NA} + w_{SA} = 1$, are introduced to assess the relevance of each coefficient in computing the global affinity value.

Clustering of $\mathrm{ODL_I}^3$ classes

The clustering phase produces an affinity tree when classes are the leaves and nodes have an associated affinity value.


University.Room
CL0
Computer_Science.Course
University.Section
CL1
0.25
0.58
0.25
tax_position_xml.ListOfStudent
CL2
Computer_Science.Office
CL3
Computer_Science.Location
CL4
0.25
0.25
0.25
University.Department
CL5
0.25
CL6
University.Research_Staff
0.57
tax_position_xml.Student
0.64
Computer_Science.CS_Person
0.7333
Computer_Science.Professor
0.685
Computer_Science.Student
0.6875
University.School_Member