SimCT: A measure of semantic similarity adapted to hierarchies of concepts

Coulibaly Kpinna Tiekoura
National Polytechnic Institute
Department of Mathematics and Computer Science
Abidjan, Ivory Coast

Brou Konan Marcellin
National polytechnic Institute
Department of Mathematics and computers science
Yamoussoukro, Ivory Coast

Achiepo Odilon
National polytechnic Institute
Abidjan, Ivory Coast

Babri Michel
National Polytechnic Institute
Abidjan, Ivory Coast

Aka Boko
University of Nangui Abrogoua
Abidjan, Ivory Coast

***Abstract*— The Calculating of the similarity between data is a key problem in several disciplines such as machine learning, information retrieval (IR) and data analysis. In some areas such as social resilience, the similarity measures can be used to find the similarities between traumatized individuals or resilience's dimensions.**

In this paper, we propose a measure of semantic similarity used in many applications including clustering and information retrieval. It relies on a knowledge base represented as a hierarchy of concepts (ontology, graph, taxonomy). Its uniqueness with respect to previous proposals is the difference between the indices of similarity that it establishes between brothers concepts located at the same hierarchical level and having the same direct ancestor. In addition, our semantic similarity measure provides better modularity in clustering compared with Wu and Palmer's similarity measure and Proxygenea 3.

Keywords- clustering, hierarchical tree, resilience, semantic similarity measure.

I. INTRODUCTION

The use of the similarity measures in a field meets a specific goal. The information retrieval, the calculation of Similarities between documents and users' queries is used to identify the relevant documents in relation to the information needs expressed by these users. In the field of clustering, these measures allow grouping objects in homogeneous classes according to their likeness.

In clustering, hierarchical representations such as ontologies are most often used to calculate the similarity between different concepts. In this proposal, we present a measure of semantic similarity to calculate the semantic proximity between the concepts of a hierarchy.

The paper is organized as follows. In Section 2, we present the properties of similarity measures followed by a state of the art of the main proposals of semantic similarity measures, in Section 3. Section 4 is devoted to the description of our proposal. Finally, in Section 5, we present our experimental results highlighting a comparison between our measure of semantic similarity and two measures, especially, the widely used measure of Wu and Palmer [1] and the ProxiGenea3 measure of Damien D. and al [2].

II. PROPERTIES OF MEASURES OF SIMILARITY AND DISSIMILARITY

A. Definition

To calculate the proximity between two objects, we can either use a similarity or a dissimilarity or a distance [3].

Similarity / dissimilarity: We call similarity or dissimilarity, any application with numerical values that quantifies the relationship between two objects, according to their point of similarity and dissimilarity. The two objects to be compared must of course be of the same type. For a similarity, the link between two individuals is stronger when its value is great. For a dissimilarity, the link is stronger when its value is small [4].

B. Properties

Positivity property: An application $d : \Omega I \times \Omega I \rightarrow \mathbb{R}$ satisfies the positivity property if and only if:

$$\forall i, j \in \Omega_I, d(i,j) \geq 0 \quad (1)$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Symmetry property: An application $d : \Omega I \times \Omega I \rightarrow \mathbb{R}$ verifies the symmetry property if and only if:

$$\forall i, j \in \Omega_I, d(i,j) = d(j,i) \quad (2)$$

Minimality property: An application $d : \Omega I \times \Omega I \rightarrow \mathbb{R}$ verifies the minimality property if and only if:

$$\forall i, j \in \Omega_I, d(i,j) = 0 \Leftrightarrow i = j \quad (3)$$

Maximality property: An application $d : \Omega I \times \Omega I \rightarrow \mathbb{R}$ checks the maximality property if and only if:

$$\forall i, j \in \Omega_I, d(i,i) \geq d(i,j) \quad (4)$$

Triangle inequality property: An application

$d : \Omega I \times \Omega I \rightarrow \mathbb{R}$ checks the triangle inequality property if and only if:

$$\forall i, j, k \in \Omega_I, d(i,j) \leq d(i,k) + d(k,j) \quad (5)$$

- The similarity is an application $s : \Omega I \times \Omega I \rightarrow \mathbb{R}+$ which verifies the properties of symmetry, of positivity and of maximality.

- The dissimilarity is an application $d : \Omega I \times \Omega I \rightarrow \mathbb{R}+$ which verifies the properties of symmetry, of positivity and of minimality.

- The Distance is an application $d : \Omega I \times \Omega I \rightarrow \mathbb{R}+$ which verifies the properties of symmetry, of positivity, of minimality and of triangle inequality.

TABLE I. SUMMARY OF PROPERTIES OF SIMILARITY INDICES / DISSIMILARITY AND DISTANCE

Measurement Type	Symmetry	positivity	minimality	maximality	triangular inequality
Similarity	X	X		X	
dissimilarity	X	X	X		
Distance	X	X	X	X	X

III. STATE OF THE ART OF SEMANTIC SIMILARITY MEASURES

The choice of a similarity relates to the type of data used. In this paper, we focus on symbolic data, specifically to structured variables. These are variables represented as a hierarchy of concepts. These variables can be single-valued or variables with values composed (qualitative variable described by several terms or quantitative variable described by a range of values) in some cases. We present in this section Semantic similarity measures most used.

The semantic similarity can be defined according to two major approaches: those that include external information to the hierarchy of concepts, for example, statistics on the use of the types of concepts [1] [5] [6] [7], and approaches based solely on the hierarchy of concepts [8] [9].

A. Similarity measures based on the hierarchy of concepts

These measures have to principle, the seeking of a taxonomic distance consisting to counting the number of edges that separate two senses. Thus, given two concepts, their similarity can be estimated from their position in the hierarchy. Every concept of the structure is represented by a node that is either a "child" called hyponym of another concept or a "parent" called hypernym.

The Measure of Rada and al. [8] is the first to use the distance between the nodes corresponding to the links of hyponymy and hypernym. The Semantic similarity between two concepts is the inverse of the distance between two concepts. More two concepts are distant, least they are similar. It is defined by:

$$Sim_{Rada}(x_i, x_j) = \frac{1}{1+dist_{edge}(x_i, x_j)} \quad (6)$$

With $dist_{edge}(x_i, x_j)$ is the length of the shortest path between two concepts x_i and x_j . According to Edge's approach, this distance can be measured by the geometric distance between the nodes representing the concepts x_i and x_j .

Leacock and Chodorow [9] are based on the fact that the lowest arcs in the hyponym's hierarchy, correspond to the smallest semantic distance. So, they have defined the following measure:

$$sim_{Lech}(x_i, x_j) = -\log \frac{longueur(x_i, x_j)}{2D} \quad (7)$$

Where D is the maximum depth or height of the hierarchy which is equivalent to the number of nodes between the highest concept in the hierarchy and the lowest concept.

$longueur(x_i, x_j)$ represents the length of the shortest path between x_i and x_j in terms of number of nodes.

The similarity measure of Wu and Palmer [1] measures the similarity between two concepts in WordNet taxonomy combining the depth of two given concepts and that of the lowest common ancestor (LCS or lowest common subsume). The authors rely on the fact that the terms are more deeply located in the taxonomy are always closer than the most general terms. It is defined as:

$$sim_{wup}(x_i, x_j) = \frac{2 \times d(LCS)}{d(x_i)+d(x_j)} \quad (8)$$

Where d (LCS) is the depth of the lowest common ancestor of the two concepts and $d(x_i)$, the depth of the concept x_i in the WordNet taxonomy.

Another measure of semantic similarity based on that of Wu and Palmer and the principle of family tree, was proposed

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

by Dudognon Gilles H. and B. Ralalason and baptized: ProxiGenea [2]. Its special feature is the inclusion of the distinction between the subsumption relations and sibling relationships unlike that proposed by Wu and Palmer. There are three versions of ProxiGenea to calculate the proximity between two concepts x_i and x_j. They are defined as follows:

$$Pg_1(x_i, x_j) = \frac{d^2(LCS(x_i,x_j))}{d(x_i)\times d(x_j)} \quad (9)$$

$$Pg_2(x_i, x_j) = \frac{d(LCS(x_i,x_j))}{d(x_i)+d(x_j)-d(LCS(x_i,x_j))} \quad (10)$$

$$Pg_3(x_i, x_j) = \frac{1}{1+d(x_i)+d(x_j)-2d(LCS(x_i,x_j))} \quad (11)$$

With $d(x_i)$ all concepts that go into the genealogy of the concept x_i from the root to x_i and $d(LCS(x_i, x_j))$, the depth of the lowest common ancestor of concepts x_i and x_j such as

$$d(LCS(x_i, x_j) = d(x_i) \cap d(x_j) \quad (12)$$

This similarity puts particular emphasis on the distance between concepts.

In [10], [11] and [12], a calculation method of the proximity between two sentences was proposed that combines a measure of structural similarity (n-gram based similarity) and the conceptual similarity measure (proxigenea 3) seen previously. The conceptual similarity between sentences p and q through an ontology is calculated as follows:

$$ss(p,q) = \frac{\sum_{x_i \in X_p} \max_{x_j \in X_p} s(x_i,x_j)}{|X_p|} \quad (13)$$

Where $s(x_i, x_j)$ is a measure of conceptual similarity between concepts x_i and x_j. That used by the authors, is ProxiGenea3 measure of Dudognon and al, presented above.

B. The similarity measures that include external information to the hierarchy of concepts

One of the most known measures of this type is that of Resnik proposed in [5] that uses the information content (IC) of the nodes (or concepts). It is generally based on a training corpus and measures the probability of finding a concept or one of his descendants in this corpus. Let X be a concept, and $p(x)$ the probability of finding X or find one of his descendants in the corpus. The information content associated with X is then defined by:

$$IC(x) = -\log(p(x)) \quad (14)$$

With $p(x) = \frac{freq(x)}{N}$ (15)

And $freq(x) = \sum_{a \in words(x)} count(a)$ (16)

Where $word(c)$ is the set of words or terms representing the concept X and concepts subsumed by X;

$freq(x)$ is the frequency of the concept in the corpus;

$count(a)$ denotes the number of occurrences of a term in the corpus;

N is the total number of occurrences of words found in the corpus.

Thus the similarity of Resnik between two concepts x_i and x_j is the following:

$$sim_{res}(x_i, x_j) = IC(LCS(x_i, x_j)) = -\log p(LCS(x_i, x_j)) \quad (17)$$

With $LCS(x_i,x_j)$ all concepts that subsume the two concepts x_i and x_j.

Another measure of semantic similarity based on the information content is proposed in [13]. Unlike the previous one, it is not based on the use of a corpus and calculates the information content of the nodes from WordNet [14] only. The hypothesis of Seco and al. is that, more a concept has descendants, the less it's informative. So, they use the hyponyms of a concept to calculate the information content thereof, as follows:

$$IC_{wn}(x) = \frac{\log\left(\frac{hypo(x)+1}{\max_{wn}}\right)}{\log\left(\frac{1}{\max_{wn}}\right)} = 1 - \frac{\log(hypo(x))+1}{\log(\max_{wn})} \quad (18)$$

Where $hypo(x)$ indicates the number of hyponyms of the concept X, and $\max_{wn}$, the number of concepts in the taxonomy.

Lin in [7] proposes a measure that is only the standardization of Resnik's measure and an extension of the measure of Wu and Palmer mentioned above. This measure reuses the concepts of information content and lowest common ancestor (LCS).

$$sim_{lin}(x_i, x_j) = \frac{2\times IC(LCS(x_i,x_j))}{IC(x_i)+IC(x_j)} = \frac{2\times\log P(LCS(x_i,x_j))}{\log P(x_i)+\log P(x_j)} \quad (19)$$

Moreover, Lin defines a measurement class of similarity based on the metric distance between two concepts, from their metric

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

distance. Thus, if the distance metric between two concepts x_i and x_j is $dist(x_i,x_j)$, their similarity is defined by:

$$sim(x_i, x_j) = \frac{1}{1+dist(x_i,x_j)} \quad (20)$$

Finally, Jiang and Conrath [6], while also based on the measurement of Resnik, proposes to calculate the similarity between two concepts as follows:

$$sim_{jc}(x_i, x_j) = \frac{1}{IC(x_i)+IC(x_2)-2IC(LCS(x_i,x_2))} \quad (21)$$

C. Other similarity measures

The following similarity measures are based on corpuses. They don't require vocabulary or grammar of the language of the text. Among them, we can cite latent semantic analysis (LSA) in [15], the explicit semantic analysis (ESA) [16] or the normalized distance from Google (Normalized Google Distance (NGD)) [17].

IV. OUR PROPOSAL sim_{CT}

Among similarity measures presented in the previous section, we are particularly interested in that of Wu and Palmer and ProxiGénéa 3 of Dudognon and al. In the hierarchy of concepts, there are two types of relationships between concepts: a sibling relationship and a subsumption relation. The sibling relationship is between two brothers-concepts and subsumption relationship is between two concepts whose meaning of one is included in the other (relationship of hyponymy and hypernymy). The measure proposed by Wu and Palmer [1] does not take sufficient account of the distinction between these relationships. It is certainly interesting, but has a limit because it essentially aims to detect the similarity between two concepts in relation to their distance from their lowest common ancestor. Dudognon and al, in their measure ProxiGenea 3, have certainly correct that aspect, however, a problem exist in the use of these two similarity measures: The concepts brothers (from the same concept father) always have the same value of similarity. In other words, for three concepts brothers data x_i , x_j and x_k , the value of similarity between the concepts x_i and x_j is the same as that between x_i and x_k . This is abnormal in our view, at practical point of view. For our part, the concepts can be from the same concept father and not have the same semantic proximity. In this section, we propose an extension of the measure ProxiGenea 3 which takes into account this difference of values of similarity between such concepts. We illustrate further, our proposal by a practical case. Through the experimental results, we compare our proposal to those of Wu and Palmer and ProxyGenea 3.

A. Notation

Either a hierarchical tree.

- Nodes x_i represent different concepts;
- $\Omega = \{x_i, x_j, ..., x_n\}$ denotes the set of all concepts of the hierarchy.
- x_{ij} means the lowest common ancestor (parent or immediate lowest common subsume (LCS)) of two concepts x_i and x_j .
- $d(x_{ij})$ is the depth of the common ancestor x_{ij} ;
- $d(x_i)$ is the depth of the hierarchy or the number of concepts which constitute his genealogy (from x_i to the root) ;
- $IC(x_i)$ is the information content of concept x_i ;

For the use of our similarity measure, we first define a metric distance between concepts. This distance is the symmetric difference between the concepts.

- $LCS(x_i, x_j)$ is all common ancestors of both concepts x_i and x_j .
- $\zeta(x_i)$ is the set of concepts that go into genealogy x_i from the root to x_i .
- $\zeta(LCS(x_i, x_j))$ is the set of concepts that have the genealogy of the common ancestor of x_i and x_j .
- $\zeta(x_i)\Delta\zeta(x_j)$ denotes the symmetric difference of $\zeta(x_i)$ and $\zeta(x_j)$ that is to say all the concepts of the two concepts genealogies x_i and x_j which are not part of their common ancestors.
- $|\zeta(x_i)\Delta\zeta(x_j)|$ denotes the number of concepts of the symmetric difference of $\zeta(x_i)$ and $\zeta(x_j)$.

$$d(x_i) = card(\zeta(x_i)) \quad (22)$$

$$d(x_{ij}) = card(LCS(x_i,x_j)) \quad (23)$$

Given $\zeta(LCS(x_i, x_j))$ that intervenes in the genealogy of each concept ($\zeta(x_i)$ and $\zeta(x_j)$), we can define the cardinal of symmetric difference as follows:

$$|\zeta(x_i)\Delta\zeta(x_j)| = card(\zeta(x_i)\Delta\zeta(x_j)) \quad (24)$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$\left|\zeta(x_i)\Delta\zeta(x_j)\right| = d(x_i) + d(x_j) - 2d(x_{ij}) \quad (25)$$

Moreover, in calculating the length of the shortest path between two concepts, we also take account of a function that measures the difference of information content between these concepts. It is defined by:

$$f_{IC}(x_i,x_j)=1-\frac{1}{2}\left|IC(x_i)-IC(x_j)\right| \quad (26)$$

With $$IC(x) = \frac{\log\left(\frac{hypo(x)+1}{\max}\right)}{\log\left(\frac{1}{\max}\right)} = 1 - \frac{\log(hypo(c))+1}{\log(\max)} \quad (27)$$

Where $hypo(x)$ indicates the number of hyponyms of concept x, and $\max$, the total number of concepts in the taxonomy.

$$IC(x_i) - IC(x_j) = \frac{\log\left(\frac{hypo(x_i)+1}{\max}\right)}{\log\left(\frac{1}{\max}\right)} - \frac{\log\left(\frac{hypo(x_j)+1}{\max}\right)}{\log\left(\frac{1}{\max}\right)} \quad (28)$$

$$IC(x_i)-IC(x_j)=\frac{\log\left(\frac{hypo(x_i)+1}{\max}\right)-\log\left(\frac{hypo(x_j)+1}{\max}\right)}{\log\left(\frac{1}{\max}\right)} \quad (29)$$

$$IC(x_i) - IC(x_j) = \frac{\log\left(\frac{hypo(x_i)+1}{hypo(x_j)+1}\right)}{\log\left(\frac{1}{\max}\right)} \quad (30)$$

Is finally obtained:

$$f_{IC}(x_i,x_j)=1-\frac{1}{2}\left|\frac{\log\left(\frac{hypo(x_i)+1}{hypo(x_j)+1}\right)}{\log\left(\frac{1}{\max}\right)}\right| \quad (31)$$

We define the length of the shortest path between two concepts x_i and x_j by:

$$dist_{CT}(x_i, x_j) = \left|\zeta(x_i)\Delta\zeta(x_j)\right| \times f_{IC}(x_i, x_j) \quad (32)$$

$$dist_{CT}(x_i, x_j) = \left(d(x_i)+d(x_j)-2d(x_{ij})\right)\left(1-\frac{1}{2}\left|IC(x_i)-IC(x_j)\right|\right) \quad (33)$$

Starting from the similarity measure of Lin [7] (20) and equation (33), we obtain our similarity measure between two concepts x_i and x_j of the hierarchy of concepts as follows:

$$sim_{CT}(x_i, x_j) = \frac{1}{1+dist_{CT}(x_i,x_j)} \quad (34)$$

That is to say:

$$sim_{CT}(x_i, x_j) = \frac{1}{1+\left(d(x_i)+d(x_j)-2d(x_{ij})\right)\left(1-\frac{1}{2}\left|IC(x_i)-IC(x_j)\right|\right)} \quad (35)$$

Or:

$$sim_{CT}(x_i,x_j)=\frac{1}{1+\left(d(x_i)+d(x_j)-2d(x_{ij})\right)\left(1-\frac{1}{2}\left|\frac{\log\left(\frac{hypo(x_i)+1}{hypo(x_j)+1}\right)}{\log\left(\frac{1}{\max}\right)}\right|\right)} \quad (36)$$

B. Checking similarity properties

- sim_{CT} verifies the symmetry property:

$\forall\ x_i, x_j \in \Omega=(x_i)_{i=1\ldots p}$

$$sim_{CT}(x_i, x_j) = \frac{1}{1+\left(d(x_i)+d(x_j)-2d(x_{ij})\right)\left(1-\frac{1}{2}\left|IC(x_i)-IC(x_j)\right|\right)}$$

$$= \frac{1}{1+\left(d(x_j)+d(x_i)-2d(x_{ji})\right)\left(1-\frac{1}{2}\left|IC(x_j)-IC(x_i)\right|\right)} = sim_{CT}(x_j, x_i)$$

- sim_{CT} verifies the property of positivity:

$\forall x_i, x_j \in \Omega=(x_i)_{i=1\ldots p}$

$d(x_i) \geq d(x_{ij})$ and $d(x_j) \geq d(x_{ij})$ from where

$$d(x_i) + d(x_j) - 2d(x_{ij}) \geq 0$$

Otherwise, $1 \geq \left|IC(x_i)-IC(x_j)\right| \quad \forall x_i, x_j \in \Omega=(x_i)_{i=1\ldots p}$

So, $1 - \frac{1}{2}\left|IC(x_i)-IC(x_j)\right| \geq 0$

The expression

$$1 + \left(d(x_i)+d(x_j)-2d(x_{ij})\right)\left(1-\frac{1}{2}\left|IC(x_i)-IC(x_j)\right|\right) \geq 0$$

Therefore

$$\frac{1}{1+\left(d(x_i)+d(x_j)-2d(x_{ij})\right)\left(1-\frac{1}{2}\left|IC(x_i)-IC(x_j)\right|\right)} \geq 0$$

$$sim_{CT} \geq 0$$

- sim_{CT} verifies the property of maximality:

The quantity $1 + \left(d(x_i)+d(x_j)-2d(x_{ij})\right)\left(1-\frac{1}{2}\left|IC(x_i)-IC(x_j)\right|\right)$ reaches its minimum with $x_i = x_j$. So,

$$1 + \left(d(x_i)+d(x_i)-2d(x_{ii})\right)\left(1-\frac{1}{2}\left|IC(x_i)-IC(x_i)\right|\right) =$$

$$1 + \left(2d(x_i)-2d(x_i)\right)\left(1-\frac{1}{2}\left|IC(x_i)-IC(x_i)\right|\right) = 1 + 0 = 1$$

$sim_{CT}(x_i, x_i) = \frac{1}{1} = 1$ = maximum of $sim_{CT}(x_i, x_j)$

So, $\forall\ x_i, x_j \in \Omega=(x_i)_{i=1\ldots p}\ sim_{CT}(x_i, x_i) \geq sim_{CT}(x_i, x_j)$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

C. Experimentation and Evaluation

We apply here, our measure sim_{CT} of conceptual similarity in a practical case and present its assessment. We first describe the experimental data and we give the experimental results from the comparison of sim_{CT} with the similarity measure of Wu and Palmer and proxiGenea 3.

1) The experimental data

For our experiment, we use here, an ontology of social resilience [18]. For the purposes of our work, we amended and supplemented this ontology (Fig. 1) by including dimensions of social resilience [19]. For the sake of readability, we present just a part of the modified ontology.

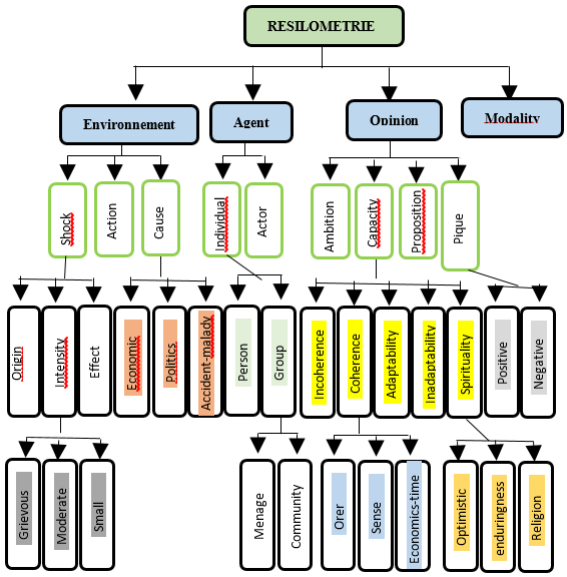

Fig. 1. Ontology of resilience processes.

We also use a simpler ontology extract (Fig. 2) similar to that used by Dudognon and al. in [2] for comparing our proposal to that of Wu and Palmer and ProxyGenea 3.

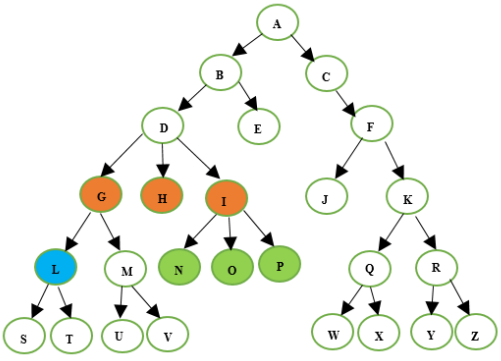

Fig. 2. Ontology extract

The objective of sim_{CT} is to distinguish the difference of similarity between concepts derived from a same hypernym or immediate parent unlike to the proposal of Wu and Palmer as well as that of Dudognon mentioned in section 3.

Two main semantic relationships characterize our ontology namely:

- Hyperonymy: relationship between a concept1 and a concept2, more general (Capacity-coherence).

- Hyponymy: relationship between a concept1 and a more specific concept2. It is the reciprocal of the hyperonymy. This relationship may be useful in information retrieval. Indeed, if one seeks all texts dealing with capacity (resilience), it may be interesting to find those who speak of consciousness, or adaptability.

The application of our similarity measure on some concepts of Fig.1, gives the following similarity values:

- Similarity of two identical concepts (shock-shock):

$$sim_{CT}(choc, choc) = \frac{1}{1+(3+3-2\times3)(1-0.5\times0)} = \frac{1}{1+0} = 1$$

- Similarity of two concepts linked by the relationship "is one" (capacity - coherence):

$$sim_{CT}(cap, coh) = \frac{1}{1+(3+4-2\times3)(1-0.5\times|-0.297|)} = 0.54$$

- Similarity of two brothers concepts (coherence-incoherence) or (coherence - spirituality):

$$sim_{CT}(coh, incoh) = \frac{1}{1+(4+4-2\times3)(1-0.5\times|-0.375|)} = 0.38$$

$$sim_{CT}(coh, spir) = \frac{1}{1+(4+4-2\times3)(1-0)} = 0.33$$

- Similarity of two concepts from two different taxonomic branches (grievous - religion):

$$sim_{CT}(grav, rel) = \frac{1}{1+(5+5-2\times1)(1-0.5\times0)} = 0.11$$

2) Comparison of conceptual similarity measures

Here, we applied the three conceptual similarity measures (simCT, proxiGenea 3 and Wu & Palmer) to the extract of simplified ontology of Fig. 2.
Table II summarizes the results obtained after application of the three measures of similarity. The aim is to show how different similarity measures take into account the different types of relationships between concepts and their relative positions in the hierarchy of concepts.

TABLE II. COMPARISON OF SEMANTIC SIMILARITY MEASURES.

N°	C_1	C_2	WP	Pg3	Sim_{CT}
1	A	B	0.67	0.67	0.85
2	A	C	0.67	0.67	0.77
3	B	D	0.67	0.5	0.79
4	B	E	0.67	0.33	0.35
5	D	G	0.8	0.5	0.56
6	D	H	0.8	0.5	0.28
7	D	I	0.8	0.5	0.43

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

8	I	N	0.85	0.5	0.30
9	I	O	0.85	0.5	0.30
10	L	S	0.88	0.5	0.30
11	B	C	0.5	0.5	0.68
12	D	E	0.5	0.33	0.32
13	G	H	0.67	0.33	0.23
14	H	I	0.67	0.33	0.20
15	N	O	0.75	0.33	0.18
16	N	P	0.75	0.33	0.18
17	O	P	0.75	0.33	0.18
18	L	M	0.75	0.33	0.26
19	S	T	0.80	0.33	0.17
20	A	A	1	1	1
21	B	B	1	1	1
22	G	K	0.16	0.16	0.29
23	L	R	0.12	0.12	0.15
24	S	Z	0.10	0.10	0.09
25	A	D	0.4	0.4	0.70
26	B	G	0.5	0.33	0.49
27	I	L	0.57	0.25	0.22
28	Q	Y	0.67	0.25	0.17
29	A	G	0.28	0.28	0.45
30	D	S	0.57	0.25	0.17

Table II presents the similarities values between concepts according to the type of relationship. Thus, lines 1 to 10 show the similarities of pairs of concepts related by subsumption relationship; Lines 11-19 show the pairs of concepts brothers ; lines 20 and 21 are the similarities values of pairs of identical concepts ; Lines 22-24 show the similarities values of concepts located on two different under taxonomic trees ; Lines 25 and 26 present the case of pairs of concepts bound by the relationship of grandfather and grand-son; Lines 27 and 28 show the similarities values of pairs of concepts linked by the uncle and nephew relationship ; Finally, lines 29 and 30 concern couples of concepts linked by the relationship of rear - grandfather and great-grand - son.

The Semantic similarity matrix of simct stemming from Table II above is:

TABLE III. MATRIX OF SEMANTIC SIMILARITY SIMCT

	A	B	C	D	E	G	H	I
A	1.00	0.52	0.53	0.35	0.50	0.29	0.40	0.31
B	0.52	1.00	0.34	0.50	0.63	0.36	0.46	0.38
C	0.53	0.34	1.00	0.25	0.34	0.20	0.27	0.22
D	0.35	0.50	0.25	1.00	0.45	0.52	0.62	0.54
E	0.50	0.63	0.34	0.45	1.00	0.32	0.25	0.29
G	0.29	0.36	0.20	0.52	0.32	1.00	0.41	0.35
H	0.40	0.46	0.27	0.62	0.25	0.41	1.00	0.38
I	0.31	0.38	0.22	0.54	0.29	0.35	0.38	1.00

The Analysis of Tables II and III allows us to observe some similarities between the three similarity measures such as the privileging subsumption relationship to sibling relationship. However, the subsumption relation is more privileged with the measures of simct and proxiGenea3 unlike that of Wu and Palmer.

All the concepts brothers which have the same number of subsumed concepts have the same Similarity value: 0.33. This is the case for example of the relationships: I-N, I-O and I-P.

Moreover, between two concepts by the sibling relationship, the closer to the father is the one who is least subsumed (which has fewer son concepts). This is the case of relationships D-G, D-H and D-I. At the level of concepts brothers, Wu and Palmer and ProxiGenea 3 gives an identical similarity value to these concepts. In practice, this is not quite realistic, in our view. Indeed, the concepts can be from the same father and not have the same characteristics. For example, in the ontology of resilience processes shown above, the "coherence" concept is semantically more similar to the concept "incoherence" than the concept "spirituality" although all from the same father. Our measure of similarity simct establishes a difference in similarity values of concepts brothers, through the calculation function of the difference in information content f_{IC} associated.

3) Evaluation of our approach

For a partition P of all the nodes in k ($k \leq n$) groups, the modularity Q is defined by:

$$Q = \sum_{C \in P} \frac{w_C}{w} - \left(\frac{D_C}{2w}\right)^2 \qquad (37)$$

Where w_C is the number of links within the class C;

D_C, the sum of the degrees of all nodes of the class C;

w , the number of links in the hierarchy of concepts.

The calculation results of modularity are presented in Table IV.

TABLE IV. COMPARISON OF MODULARITY BASED ON THE NUMBER OF CLASSES.

Number of classes	Modularity		
	WP	PG	Simct
2	0.131	0.154	0.156
3	0.129	0.129	0.129
4	0.145	0.150	0.160

Looking at the graph in Fig. 3 below, we notice a superiority of our proposal in terms of clustering quality than that of Wu and Palmer and proxiGénéa3.

Fig. 3. Comparison of modularities

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

V. CONCLUSION AND OUTLOOK

In this paper, we proposed a conceptual semantic similarity measure on taxonomic data type. It can be used in many applications including automatic classification or for information retrieval (IR). Our proposal, compared to the Wu and Palmer measure and ProxiGenea 3 provides a better quality of clustering and establishes a difference in the semantic similarity between concepts brothers from the same father.

Our future work will consist initially, to apply our similarity measure on larger data to confirm the results presented in this article in a broader context. Secondly, we will consider extending this measure to other types of symbolic data.

REFERENCES

[1] WU, Z. et PALMER, M., Verb semantics and lexical selection, Proceedings of the 23rd Annual Meetings of the Associations for Computational Linguistics, p. 133-138, 1994

[2] Dudognon, D., Hubert, G., & Ralalason, B. J. V., Proxigénéa: Une mesure de similarité conceptuelle. In Proceedings of the Colloque Veille Stratégique Scientifique et Technologique (VSST 2010), 2010, October.

[3] Bisson G. La similarité: une notion symbolique/numérique, In Apprentissage symbolique - numérique, éd. par Moulet B. Editions CEPADUES, pp. 169 - 201, 2000.

[4] Celeux G., Diday E., Lechevallier Y., Govaert G. and Ralambondrainy H. Classification automatique des données, Editions Dunod, Paris, 1989.

[5] RESNIK, P., Using information content to evaluate semantic similarity in a taxonomy, IJCAI, p. 448-453, 1995.

[6] JIANG, J. et CONRATH, D.W., Semantic Similarity based on Corpus Statistics and Lexical Taxonomy, Proceedings of the International Conference on Research in Computational Linguistics (ROCLING), Taiwan, 1997.

[7] LIN, D., An information-theoric definition of similarity, Proceedings of the 15th international conference on Machine Learning, p. 296-304, 1998.

[8] RADA, R., MILI, H., BICKNELL, E. et BLETTNER, M., Development and application of a metric on semanticnets. Systems, Man and Cybernetics, IEEE Transactions on, 19(1): p. 17-30, 1989.

[9] LEACOCK, C., MILLER, G. A., et CHODOROW, M., Using corpus statistics and WordNet relations for sense identification, Comput. Linguist. 24, 1, 147-165, 1998

[10] Buscaldi, D., Flores, J. J. G., Meza, I. V., & Rodrıguez, I., SOPA: Random Forests Regression for the Semantic Textual Similarity task.SemEval-2015, 132, 2015.

[11] Buscaldi, D., Le Roux, J., Flores, J. J. G., & Popescu, A., Lipn-core: Semantic text similarity using n-grams, wordnet, syntactic analysis, esa and information retrieval based features. In Second Joint Conference on Lexical and Computational Semantics (p. 63), 2013, June.

[12] Buscaldi, D., Tournier, R., Aussenac-Gilles, N., & Mothe, J., Irit: Textual similarity combining conceptual similarity with an n-gram comparison method. In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (pp. 552-556). Association for Computational Linguistics, 2012, June.

[13] Seco, N., Veale, T., & Hayes, J., An intrinsic information content metric for semantic similarity in WordNet. In ECAI (Vol. 16, p. 1089), 2004, August.

[14] Fellbaum, C., A semantic network of english: the mother of all WordNets. In EuroWordNet: A multilingual database with lexical semantic networks (pp. 137-148). Springer Netherlands, 1998.

[15] Deerwester, S., Dumais, S.T., Furnas, G. W., Landauer, T.K., and Harshman, R., Indexing by latent semantic analysis. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 41(6) : 391-407, 2010, October.

[16] Gabrilovich, E. and Markovitch, S., Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In In Proceedings of the 20th International Joint Conference on Artificial Intelligence, pages 1606-1611, 2007.

[17] Cilibrasi, R. L. and Vitanyi, P.M.B. , The google similarity distance. IEEE Trans. On knowl. And Data Eng., 19(3) :370-383, 2007.

[18] ACHIEPO OdilonYapo M., Les bases fondamentales de la Résilométrie, une science de modélisation de la souffrance. Journée Scientifique « Café Résilience », Février 2015.

[19] COULIBALY Kpinna Tiekoura, Odilon Yapo M. ACHIEPO, Brou Konan Marcellin, Michel Babri. « Resilometrical modeling of interactions in social resilience dimensions ». International Journal of Computer Science Issues (IJCSI), Volume 12, Issue 4, July 2015.

[20] Newman M.E.J., Girman M., Finding an evaluating community structure in networks.Physical Review E, 69(6), 2004.

[21] Van Dongen S.M., Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, 2000.

AUTHORS PROFILE

Coulibaly Kpinna Tiekoura is now PhD student at the Department of Mathematics and Computer Science of National Polytechnic Institute (Yamoussoukro, Ivory Coast). He received his M.S. degree in data processing from University of Nangui Abrogoa in 2013. He is a member of the international resilience research group (UMI Resilience, IRD) and is also a member of the Laboratory of Computer Sciences and Telecommunications (INP-HB) Abidjan, Ivory Coast. His research interests include the mathematical modeling, the resilience process, Multidimensional Statistics, Artificial Intelligence, Machine Learning, Data Mining and Data Science. His works are centered on clustering methods adapted to resilience process.

Brou Konan Marcellin is a Ph-D in Computer Science and Teacher researcher at the Houphouët Boigny National Polytechnic Institute (INP-HB) of Yamoussoukro (Ivory Coast). He is the Director of the Department of Mathematics and Computer Science. He is a Member of Laboratory in Computer Sciences and Telecommunications (INPHB). His interests are information systems, database and programming languages.

Odilon Yapo M. ACHIEPO is a statistician-economist Engineer (ENSEA Abidjan, Ivory Coast) and has a Master degree in Computer Science with specialization in Artificial Intelligence and Business Intelligence (Polytechnic School of Nantes, France). He is a Ph-D student in Mathematics and Information Technologies (EDP-INPHB Yamoussoukro, Ivory Coast). He is also a Teacher-researcher in University of Korhogo (Ivory Coast), International Senior-Expert Consultant, member of the international resilience research group (UMI Resilience, IRD) and is a member of the Laboratory of Computer Sciences and Telecommunications (INP-HB) Abidjan, Ivory Coast. His centers of interests are Computational Mathematics, Multidimensional Statistics, Artificial Intelligence, Machine Learning, Data Mining and Data Science. He also is the author-creator of the Resilometrics, a modeling discipline which consists on developing and applies computational models for measure, analyze and simulate social resilience process.

Babri Michel, PhD, is now Senior Lecturer in data processing. He teaches data processing and telecommunication in INP-HB. He is the Director of the Laboratory of Computer Sciences and Telecommunications (INP-HB) Abidjan, Ivory Coast. The topics of his current interest of research include distributed networks, cloud computing, convergent mobile networks, big

Data and software defined networks.

Aka Boko, is a titular professor in computer sciences and physical sciences at Nangui Abrogoua University (Abidjan, Ivory Coast). He is the Director of the Department of Mathematics and Computer Sciences at that University. His interests are information systems, big data, programming languages and Artificial Intelligence.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500