*Intelligent Information Management*, 2010, 2, 65-79
doi:10.4236/iim.2010.22009 Published Online February 2010 (http://www.scirp.org/journal/iim)

# Prototypicality Gradient and Similarity Measure: A Semiotic-Based Approach Dedicated to Ontology Personalization

**Xavier Aimé[1,3], Frédéric Furst [2], Pascale Kuntz[3], Francky Trichet[3]**
[1]*Société TENNAXIA, Paris, France*
[2]*MIS - Laboratoire Modélisation, Information et SystèmeUniversity of Amiens, Amiens, France*
[3]*LINA - Laboratoire d'Informatique de Nantes Atlantique* (*UMR-CNRS* 6241)*, University of Nantes, Team "Knowledge and Decision", Nantes, France*
*Email*: *xaime@tennaxia.com, frederic.furst@u-picardie.fr, {pascale.kuntz,francky.trichet}@univ-nantes.fr*

## Abstract

This paper introduces a new approach dedicated to the Ontology Personalization. Inspired by works in Cognitive Psychology, our work is based on a process which aims at capturing the user-sensitive relevance of the *categorization process*, that is the one which is really perceived by the end-user. Practically, this process consists in decorating the Specialization/Generalization links (*i.e.* the is-a links) of the hierarchy of concepts with 2 gradients. The goal of the first gradient, called Conceptual Prototypicality Gradient, is to capture the user-sensitive relevance of the categorization process, that is the one which is perceived by the end-user. As this gradient is defined according to the three aspects of the semiotic triangle (*i.e.* intentional, extensional and expressional dimension), we call it Semiotic based Prototypicality Gradient. The objective of the second gradient, called Lexical Prototypicality Gradient, is to capture the user-sensitive relevance of the lexicalization process, *i.e.* the definition of a set of terms used to denote a concept. These gradients enrich the initial formal semantics of an ontology by adding a pragmatics defined according to a context of use which depends on parameters like culture, educational background and/or emotional context of the end-user. This paper also introduces a new similarity measure also defined in the context of a semiotic-based approach. The first originality of this measure, called SEMIOSEM, is to consider the three semiotic dimensions of the conceptualization underlying an ontology. Thus, SEMIOSEM aims at aggregating and improving existing extensional-based and intentional-based measures. The second originality of this measure is to be context-sensitive, and in particular user-sensitive. This makes SEMIOSEM more flexible, more robust and more close to the end-user's judgment than the other similarity measures which are usually only based on one aspect of a conceptualization and never take the end-user's perceptions and purposes into account.

**Keywords:** Semantic Measure, Conceptual Prototypicality, Lexical Prototypicality, Gradient, Ontology Personalization, Semiotics

## 1. Introduction

This paper deals with Subjective knowledge, that is knowledge which is included in the semantic and episodic memory of Human Being [1]. Objective knowledge, which can be expressed through textual, graphic or sound documents, corresponds to what must be captured within a Domain Ontology, as it is specified by the consensual definition of T. Gruber [2]: "an ontology is a formal and explicit specification of a shared conceptualization". The advent of the Semantic Web and the standardization of a Web Ontology Language (OWL) have led to the definition and the sharing of a lot of ontologies dedicated to scientific or technical fields. However, with the current emergence of Cognitive Sciences and the development of Knowledge Management applications in Social and Human Sciences, subjective knowledge becomes an unavoidable subject and a real challenge, which must be integrated and developed in Semantic Web, and more generally in Ontology Engineering. Our work aims at

 

providing measures dedicated to the personalization of a Domain Ontology (which by definition only includes Objective knowledge). This personalization process mainly consists in adapting the content of an ontology to its context of use. This latter usually implies an end-user and therefore mobilizes *Subjective knowledge* which is defined according to several parameters such as culture, educational background and emotional state. Our approach of ontology personalization aims at taking these parameters into account in order to reflect the relevance users of ontologies perceive on the is-a hierarchies and to what extent the terms associated to the concepts are representative.

Our work is based on the semiotic triangle, as defined by the linguists Ogden and Richard [3]. The three corners of this triangle are 1) the reference (*i.e.* the intentional dimension) which is an unit of thought defined from abstraction of properties common to a set of objects (*i.e.* the concepts of an ontology and their properties), 2) the referent (*i.e.* the extensional dimension) which corresponds to any part of the perceivable or conceivable world (*i.e.* the instances of concepts) and 3) the term (*i.e.* the expressional dimension) which is a designation of an unit of thought in a specific language (*i.e.* the linguistic expressions used to denote the concepts). The goal of our first gradient is to capture the user-sensitive relevance of the categorization process, that is the one which is perceived by the end-user. As this gradient is defined according to the three aspects of the semiotic triangle (*i.e.* intentional, extensional and expressional dimension), we call it *Semiotic-based Prototypicality Gradient*. The goal of our second gradient, called *Lexical Prototypicality Gradient*, is to capture the user-sensitive relevance of the lexicalization process, *i.e.* the definition of a set of terms used to denote a concept.

This paper also introduces a new similarity measure also defined in the context of a semiotic-based approach. The first originality of this measure, called SEMIOSEM, is to consider the three dimensions of the conceptualization underlying an ontology: the intention (*i.e.* the properties used to define the concepts), the extension (*i.e.* the instances of the concepts) and the expression (*i.e.* the terms used to denote both the concepts and the instances). Thus, SEMIOSEM aims at aggregating and improving existing extensional-based and intentional-based measures, with an original expressional one. The second originality of this measure is to be context-sensitive, and in particular user-sensitive. Indeed, SEMIOSEM exploits multiple information sources: 1) a textual corpus, validated by the end-user, which must reflect the domain underlying the ontology which is considered, 2) a set of instances known by the end-user, 3) an ontology enriched with the perception of the end-user on how each property associated to a concept c is important for defining c and 4) the emotional state of the end-user. The importance of each source can be modulated according to the context of use and SEMIOSEM remains valid even if one of the sources is missing. This makes our measure more flexible, more robust and more closer to the end-user's judgment than the other similarity measures.

The rest of this paper is structured as follows. Section 2 presents the formal definition of the Semiotic-based Prototypicality Gradient (SPG) which corresponds to the aggregation of three components related to the *intentional*, the *expressional* and the *extensional* dimension of a conceptualization. Section 3 introduces the formal definition of the Lexical Prototypicality Gradient (LPG). Section 4 briefly introduces some well-known similarity measures and describes in detail SEMIOSEM: the basic foundations, the formal definitions, the parameters of the end-user and their interactions. Section 5 presents the tool TOOPRAG which implements our approach of ontology personalization; it also introduces some experimental results defined in two contexts: 1) a case study dedicated to the analysis of texts describing the Common Agricultural Policy (CAP) of the European Union and 2) a case study dedicated to Legal Intelligence within regulatory documents related to the domain "Health, Safety and Environment" (HSE).

## 2. Semiotic-Based Prototypicality Gradients (SPG)

Defining an ontology O of a domain D at a precise time T consists in establishing a consensual synthesis of individual knowledge belonging to a specific endogroup; an endogroup is a set of individuals which share the same distinctive signs and, therefore, characterize a community. For the same domain, several ontologies can be defined by different endogroups. We call *Vernacular Domain Ontologies* (VDO) this kind of resources[1]. This property is also described by E. Rosch as *ecological* [4, 5], in the sense that although an ontology belongs to an endogroup, it also depends on the context in which it evolves. Thus, given a domain D, an endogroup G and a time T, a VDO depends on three factors, characterizing a precise context: 1) the culture of G, 2) the educational background of G and 3) the emotional state of G. In this way, a VDO can be associated to a pragmatic dimension. Indeed, a same VDO can be viewed (and used) from multiple points of view, where each point of view, although not reconsidering the formal semantics of D, allows us to adapt 1) the degrees of truth of the is-a links defined between concepts and 2) the degrees of expressivity of the terms used to denote the concepts. We call *Personalized Vernacular Domain Ontologies* (PVDO) this kind of resources. Our work is based on the funda-

[1] *Vernacular,* which comes from the latin word *vernaculus*, means native For instance, vernacular architecture, which is based on methods of building which use locally available resources to address local needs, tends to evolve over time to reflect the environmental, cultural and historical context in which it exists.

 

mental idea that all the sub-concepts of a decomposition are not *equidistant* members, and that some sub-concepts are more representative of the super-concept than others. This phenomenon is also applicable to the set of terms used to denote a concept. This assumption is validated by works in Cognitive Psychology [1,6]. Formally, our gradient is based on a Vernacular Domain Ontology (VDO), given a field D and an endogroup G. This type of ontology is defined by the following tuple: $O_{(D,G)}$= {*C, P, I,* $\Omega_{(D,G)}$ , $\leq^C$, $\sigma^P$,*L* } where:

- *C, P, I* represent respectively the disjoined sets of concepts, properties[2] and instances;
- $\Omega_{(D,G)}$ is a set of documents (e.g. text, graphic or sound documents) related to a domain *D* and shared by the members of the endogroup *G*;
- $\leq^C$: $C \times C$ a partial order on *C* defining the hierarchy of concepts ($\leq^C(c_1,c_2)$ means that the concept $c1$ subsumes the concept $c_2$);
- $\sigma^P$: P $\rightarrow C \times C$ defines the domain and the range of a property;
- $L = \{L_C, ftermc\}$ is the lexicon related to the dialect of G where a) *LC* represents the set of terms associated to C, b) the function $ftermc : C \rightarrow (L_C)^n$ which returns the tuple of terms used to denote a concept.

We define $spg_{G;D} : C \times C \rightarrow [0; 1]$ the function which, for all pairs of concepts $c_f$, $c_p \in$ C such as it exists an is a link between the super-concept $c_p$ and the sub-concept $c_f$, returns a real (null or positive value) which represents the conceptual prototypicality gradient of this link, in the context of a PVDO dedicated to a domain D and an endogroup G. For two concepts $c_p$ and $c_f$, this function is formally defined as follows:

$$spgG,D(c_p,c_f) = [\alpha * intentional(c_f,c_p) + \beta * expressional_{G,D}(c_f,c_p) + \gamma * extensional_{G,D}(c_f,c_p)]^{\delta} \quad (1)$$

with 1) $\alpha + \beta + \gamma = 1$, where $\alpha \geq 0$ is a weighting of the intentional component, $\beta \geq 0$ is a weighting of the expressional component, $\gamma \geq 0$ is a weighting of the extensional component, and 2) $\delta \geq 0$ is a weighting of the mental state of the endogroup G. Our gradient is based on the three dimensions introduced by Morris and Peirce in their theory of semiotics [7]: 1) the signified, *i.e.* the concept defined in intention, 2) the referent, *i.e.* the concept defined in extension via its instances, and 3) the signifier, *i.e.* the terms used to denote the concept. The main advantage of our approach is that 1) it integrates both the intentional, extensional and expressional dimension of a conceptualization for defining how a sub-concept is *representative*/*typical* of its super-concept and the influence of these dimensions can be modulated via the α, β and γ parameters and 2) it allows us to modulate this representativeness according to an emotional dimension via the $\delta$ parameter. The values of $\alpha$, $\beta$ and $\gamma$ are defined manually according to the context of the ontology personalization process. Indeed, when no instances (or few) are associated to the ontology then it is relevant to minimize the influence of the extensional dimension by assigning a low value to $\gamma$. In a similar way, when the ontology does not include properties then it is relevant to minimize the influence of the intentional dimension by assigning a high value to $\alpha$. And when the ontology is associated to a huge and rich textual corpora then it is relevant to maximize the influence of the expressional dimension by assigning a high value to $\beta$ given that $\alpha + \beta + \gamma = 1$. The value of $\delta$ is used to modulate the influence of the emotion state on the perception of the conceptualization. Multiple works on the influence of emotions on human evaluation have been done in psychology [8, 9]. The conclusion of these works can be summarized as follows: when we are in a negative mental state (e.g. fear or nervous breakdown), we tend to focus us on what appears to be the more important from an emotional point of view. In our context, this consists in reducing the universe to what is very familiar; for instance, our personal dog (or the one of a neighbor) which at the beginning is inevitably the most characteristic of the category becomes the quasi unique and quasi unique dog. Respectively, in a positive mental state (e.g. love or joy), we are more open in our judgment and we accept more easily the elements which are not yet be considered as so characteristic. According to [10], a *negative* mental state leads to the reduction of the value of representation, and conversely for a *positive* mental state. Thus, we characterize: 1) a *negative* mental state by a value $\delta \in ]1,+\infty[$, 2) a *positive* mental state by a value $\delta \in ]0,1[$, and 3) a *neutral* mental state by the value 1. When the value of δ is low, the value of the gradients associated to the concepts which are initially not considered as being so representative increases considerably, because a positive state favours openmindedness, self-actualization, etc. Conversely, when the value of δ is high (*i.e.* a strongly negative mental state), the effect is to only *select* the concepts which own a high value of typicality, eliminating *de facto* the other concepts.

[2] Properties include both attributes of concepts and domain relations.

## 2.1. Intentional Component

The intentional component of our gradient aims at taking 1) the structure of a conceptualization and 2) the intentional definition of its components into account. In order to compare two concepts from an intentional point of view, we propose a measure based on the properties shared by the sub-concepts as developed in [11,12]. For each concept $c \in C$, we define a *Characteristic Vector* (CV) vc= ($v_{c1}$ , $v_{c2}$ , … , $v_{cn}$) with $n = |P|$, and $v_{ci} \in [0, 1]$, $\forall i \in [1 , n]$ a weight assigned to each

 

property $p_i$ of $P$. A concept is defined by the union of all the properties whose weight is not null. The set of concepts corresponds to a point cloud defined in a space with $|P|$ dimensions. Weights associated with the properties have to satisfy a constraint related to the *isa* relationship: a concept $c$ is subsumed by a concept $d$ (noted $\leq^C (d, c)$) if and only if $v_{ci} \geq v_{di}$, $\forall i \in [1, n]$ with $n = |P|$. For any $c \in C$, we define a prototype concept from all the sub-concepts of c. This prototype concept is characterized by a *Prototype Vector* (PV) tc= ($t_{c1}$ , $t_{c2}$ , … , $t_{cn}$). Prototype concepts correspond to summaries of semantic features characterizing categories of concepts. They are stored in the episodic memory, and are used in the process of categorization per comparison. In our work, we consider that a prototype concept of a concept $c$ corresponds to the barycenter of the point cloud formed by the set of the CV of all concepts belonging to the descent of $c$[3]. Thus, the Prototype Vector of a concept $c$ is formally defined as follows:

$$\vec{t}_c = \frac{1}{\sum_{s \in S} \lambda(s)} \sum_{s \in S} \lambda(s) \vec{v}_s \quad (2)$$

where:

- $\lambda$ ($s$) is equal to ($depthtree(c) - depth(s) + 1$) / $depthtree(c)$ with 1) $depthtree(c)$, the depth of the sub-tree having for root $c$ and 2) $depth(s)$, the depth of $s$ in the sub-tree having for root $c$;
- $S$, the set of concepts belonging to the descent of $c$.

The objective of the coefficient $\lambda$ ($s$) (for a concept $s$) is to relativize the properties which are hierarchically distant from the super-concept (cf. the use of the ratio of depths)[4]. Note that if we consider only the sub-concepts of $c$ (*i.e.* only one level of hierarchy), then $\lambda(s) = 1$, $\forall s \in S$ and we find the formula of the PV defined by [12]:

$$\vec{t}_c = \frac{1}{|S|} \sum_{s \in S} \vec{v}_s \quad (3)$$

In our work, we advocate the following principle: the more a concept is close to the prototype concept, the more it is representative of its super-concept. We consider this value as being the Euclidean normalized distance between 1) the PV of the super-concept and 2) the CV of the subconcept which is considered; it corresponds to the normalized distance between a point and the barycenter of the point cloud. The function *intentional*: $C \times C \rightarrow [0; 1]$ is formally defined as follows:

$$intentional(c_f, c_p) = 1 - dist(\vec{t}_{cp}, \vec{v}_{cf}) \quad (4)$$

The more the value of this function is near to 1, the more the concept $c_f$ is *representative/typical* of the concept $c_p$, from an intentional point of view.

## 2.2. Expressional Component

The expressional component of our gradient aims at taking the expressional view of a conceptualization into account, through the terms used to denote the concepts. This approach is based on the observation frequency of a concept related to a domain D, in an universe of the endogroup G. In this way, the more an element is frequent in the universe, the more it is considered as *representative/typical* of its category. This notion of typicality is introduced in the work of E. Rosch [4,5]. In our context, the universe of an endogroup is composed of the set of documents identified by $\Omega_{(D,G)}$. Our approach is inspired by the idea of Information Content introduced by Resnik [13]. Indeed, this is not because an idea is often expressed that it is really true and objective. Psychologically, it is acknowledged that the more an event is presented (in a frequent way), the more it is judged probable without being really true for an individual or an endogroup; this is one of the ideas supported by A. Tversky in its work on the evaluation of uncertainty [14]. The function *expressional*: $C \times C \rightarrow [0, 1]$ is formally defined as follows[5]:

$$expressional_{G,D}(c_f, c_p) = \frac{Info(c_f)}{Info(c_p)} \quad (5)$$

where :

$$Info(c_f) = \sum_{term \in world(c)} \left( \frac{count(term)}{N} * \frac{count(doc, term)}{count(doc)} \right) \quad (6)$$

with:

- $Info(c)$ defines the information content of the concept $c$;
- $count(term)$ returns the weighting number of term occurrences in the documents of $\Omega_{(D,G)}$. Note that this function takes the structure of the documents into account. Indeed, in the context of a scientific article, an occurrence of a term $t$ located in the keywords section is more important than another occurrence of $t$ located in the summary or the body of text. Thus, this function is formally defined as follows:

$$count(term) = \sum_{i=1}^{m} M_{term,i} \quad (7)$$

[3]We understand by descent all the sub-concepts of $c$, from generation 1 to n (*i.e.* the leaves).

[4]Contrary to [12], we propose to extend the calculation of the prototype to all the descent of a concept, and not only to its direct sub-concepts (*i.e.* only one level of hierarchy). Indeed, we think that all the concepts belonging to the descent (and in particular the leaves) contribute to the definition of the prototype from a cognitive point of view.

[5]This function is only applicable if it exists 1) a direct *is a* link between the super-concept $c_p$ and the sub-concept $c_f$, with an order relation $c_f \leq c_p$, or 2) an indirect link composed of a serie of *is a* links between the $c_p$ and $c_f$.

 

where $M_{term,i} \in Z$ is the hierarchical coefficient relating to the position (in the structure of the document) of the ith occurrence of the term. The values of these coefficients are fixed in a manual and consensual way by the members of the endogroup.

- *count*(*doc,term*) returns the number of documents of $\Omega_{(D,G)}$ where the term appears;
- *count*(*doc*) returns the number of documents of $\Omega_{(D,G)}$;
- *world*(*c*) returns all the terms concerning the concept *c* via the function $f_{termC}$ and all its sub-concepts from generation 1 to generation *n*;
- *N* is the sum of all the weighting numbers of occurrence of all the terms contained in $\Omega_{(D,G)}$.

Intuitively, the function *Info*(*c*) measures "the ratio of use" of a concept in an universe, by using first the terms directly associated to the concept and then, by using the terms associated to all its sub-concepts, from generation 1 to generation *n*. We balance each frequency by the ratio between the number of documents where the term is present and the global number of documents. An idea which is frequently presented in few documents is less relevant that an idea which is perhaps less supported in each document but which is presented in a lot of documents of the endogroup's universe.

### 2.3. Extensional Component

The extensional component of our gradient aims at taking the extensional view of a conceptualization into account, through the instances. This approach is based on the quantity of instances of a concept related to a domain D, in an universe of the endogroup G. In this way, the more a concept is frequent in the universe (because it owns a lot of instances), the more it is considered as representative/typical of its category. The function *extensional*: $C \times C \rightarrow [0, 1]$ is formally defined as follows:

$$extensional_{G,D}(c_f, c_p) = \frac{1}{1 - \log(\frac{count_I(c_f)}{count_I(c_p)})} \qquad (8)$$

Where the function $count_I(c)$: $C \times I \rightarrow Z$ return the number of instances $i \in I$ of a concept $c \in C$. The form ($1/1-log(x)$) has been adopted in order to obtain a non-linear behavior which is more close to human judgment.

## 3. Lexical Prototypicality Gradient (LPG)

The goal of the Lexical Prototypicality Gradient (LPG) is to evaluate the fact that the terms used to denote a concept have not the same representativeness within the endogroup. Indeed, the question is the following: "why do we more frequently name the concept x with the term y rather than z?" To define these lexical variations, we propose to adapt the expressional component of our conceptual gradient previously defined Thus, our formula is based on the Information Content of a concept, by using the ratio between the frequency of use of the term and the sum of the appearance frequencies of all the terms related to the concept in $\Omega_{(D,G)}$. We define define $lpg_{G,D} : L_C \times C \rightarrow [0,1]$ the function, which for all concept $c \in C$ and the term $t \in L_C$ such as $t \in f_{termC}(c)$, returns a positive or null value representing the lexical prototypicality gradient of this term, and this for a domain D and an endogroup G. This function is formally defined as follows:

$$lpg_{G,D}(t,c) = \frac{1}{1 - \log(\frac{count(t)}{\sum count(f_{termC}(c))})} \qquad (9)$$

## 4. SEMIOSEM: A Semiotic-Based Similarity Measure

Currently, the notion of similarity has been highlighted in many activities related to Ontology Engineering such as ontology learning, ontology matching or ontology population. In the last few years, a lot of measures for defining concept (dis-)similarity have been proposed. These measures can be classified according to two approaches: 1) extensional-based measures such as [13,15–16] or [17] and 2) intentional based measures such as [18,19] or [20]. Most of these measures only focus on one aspect of the conceptualization underlying an ontology, mainly the intention through the structure of the subsumption hierarchy or the extension through the instances of the concepts or the occurences of the concepts in a corpus. Moreover, they are usually sensitive to the structure of the subsumption hierarchy (because of the use of the more specific common subsumer) and, therefore, they are dependent on the modeling choices. Finally, these measures never consider the end-user's perceptions of the domain which is considered [21]. Our goal is to provide a measure more flexible, more robust and more closer to the end-user's judgment than the other similarity measures.

### 4.1. Current Intentional-Based Similarity Measures

Intentional-based measures are founded on the analysis of the structure of a semantic network. In the field of ontology engineering, the hierarchy of concepts is considered as a directed graph (where the nodes correspond to the concepts and the edges correspond to taxonomic links). Intuitively, these works are based on the following principle: concept *A* is more similar to concept *B* than concept *C*, if the distance from *A* to *B* (in the graph) is shorter than the one from *A* to *C*. This distance can be

 

calculated following different ways. [18] considers this distance, noted $distedge(c_1, c_2)$, as being the length of the shortest path between two concepts. The similarity between $c_1$ and $c_2$ is defined as follows:

$$Sim_{Rad}(c_1, c_2) = \frac{1}{dist_{edge}(c_1, c_2)} \quad (10)$$

[13] enhances this definition by introducing the maximum depth of the hierarchy, noted *max*.

$$Sim_{Res}(c_1, c_2) = \frac{2 * max}{dist_{edge}(c_1, c_2)} \quad (11)$$

[19] standardizes this latter measure in order to obtain results in the interval [0, 1].

$$Sim_{Lea}(c_1, c_2) = -\log(\frac{dist_{edge}(c_1, c_2)}{2 * max}) \quad (12)$$

[20] suggests another measure which takes the depth of the concepts into account. The similarity between $c_1$ and $c_2$, with $depth(c_i)$ the depth of the concept $c_i$ in the hierarchy and $c$ the Most Specific Common Subsumer (MSCS) of $c_1$ and $c_2$, is defined as follows:

$$Sim_{Wu}(c_1, c_2) = \frac{2 * depth(c)}{depth(c_1) + depth(c_2)} \quad (13)$$

These measures only consider the is-a links between concepts and not the semantics of concepts. Thus, they can be incorrect (concepts with high similarity can be semantically far from each other) or incomplete (for concepts semantically similar but not very close in the hierarchy, the measure can be very low). Another intentionalbased approach for defining a similarity measure consists in analyzing and comparing the properties of the concepts. For illustration, let us consider two objects on which we can sit down: an armchair and a chair. These two objects share common properties and there are other properties which differentiate them without ambiguity. From the comparison of these properties, it is possible to state that a chair is more close to an armchair than a stool. Indeed, in set theory, we can state that two concepts are close if the number of common properties is greater than the number of distinct properties. [14] suggests the following function:

$$Sim_{Tversky}(c_1, c_2) = \alpha.comm(c_1, c_2) - \beta.diff(c_1, c_2) - \lambda.diff(c_2, c_1) \quad (14)$$

where α, β, γ.are constants.

## 4.2. Current Extensional-Based Similarity Measures

Extensional-based measures have first been inspired by the measure of Jaccard [22]:

$$Sim_{Jaccard}(c_1, c_2) = \frac{|Ic_1 \cap Ic_2|}{|Ic_1| + |Ic_2| - (|Ic_1| \cap |Ic_2|)} \quad (15)$$

where $|Ic|$ is the number of instances of the concept $c$.

According to [17], this approach is not really appropriate to ontologies because two concepts can be similar without having common instances. [17] proposes a new measure which does not evaluate the extension overlap but the variation of the cardinality of the extensions of the considered concepts w.r.t. the cardinality of the extension of their *MSCS*.

$$Sim_{Ama}(c_1, c_2) = \frac{min(|Ic_1|, |Ic_2|)}{|I_{gcs(c_1, c_2)}|}(1 - \frac{|I_{gcs(c_1, c_2)}|}{|I|})(1 - \frac{min(|Ic_1|, |Ic_2|)}{|I_{gcs(c_1, c_2)}|}) \quad (16)$$

where $gcs(c_1, c_2)$ is the *MSCS* of $c_1$ and $c_2$, $|Ic|$ is the number of instances of the concept c, and $|I|$ is the total number of instances of the ontology.

Most of the current extensional-based measures are founded on the notion of "Information Content" (IC) introduced by Resnik [13]. The main idea consists in measuring the similarity of concepts on the ground on the amount of information that they share. [13] approximates the IC of a concept $c$ to the probability $p(c)$ to have occurences of $c$ in a given corpus. Thus, IC is defined as follows:

$$\psi(c) = -log(p(c)) \quad (17)$$

where :

$$p(c) = \frac{\sum_{n \in word(c)} count(n)}{N}$$

$N$ represents the number of the occurrences of the terms of all the concepts in the corpus, *words(c)* represents the set of terms used to denote the concept c and all its sub-concepts. This measure assumes that each term is associated to one and only one concept. [23] deals with this problem by modifying the calculation of the appearance frequency of a term by:

$$p(c) = \frac{\sum_{n \in word(c)} \frac{count(n)}{nb_{classe}(n)}}{N} \quad (18)$$

The similarity measure advocated by [13] is based on the most informative common subsumer of $c_1$, $c_2$ (*i.e.* the common subsumer which owns the most important IC; this is not necessary the MSCS of $c_1$, $c_2$). The similarity between $c_1$ and $c_2$, where $S(c_1, c_2)$ is the set of concepts which subsume both $c_1$ and $c_2$, is defined as follows:

$$Sim_{Res2}(c_1, c_2) = max_{c \in S(c_1, c_2)} \psi(c) \quad (19)$$

[15] suggests another measure based on the common IC of the concepts. The similarity between $c_1$ and $c_2$, with *ppc* the concept in $S(c_1; c_2)$ which minimizes $\psi(c)$, is

 

defined as follows:

$$Sim_{Lin}(c_1, c_2) = \frac{2 * \psi(ppc)}{\psi(c_1) + \psi(c_2)} \quad (20)$$

Based on this approach of IC, [16] advocates the following measure:

$$Sim_{Jiang}(c_1, c_2) = \sum_{c \in path(c_1, c_2) - MSCS(c_1, c_2)} [\psi(c) - \psi(SC(c))] * TC(c, SC(c)) \quad (21)$$

where $TC(c_i, c_j)$ is a weighting of the edge connecting $c_i$ to $c_j$ such as $c_j = SC(c_i)$, and $SC(c)$ is the super-concept of $c$.

### 4.3. Semiosem: A Semiotic-Based Similarity Measure

SEMIOSEM is a conceptual similarity measure defined in the context of a PVDO; it takes as input a PVDO and the three additional resources:

- a set of **instances** supposed to be representative of the end-user's conceptualization (for instance, in the case of a business information system, these instances are the customers the end-user deals with, the products he/she sells to them, etc);
- a **corpus** given by the end-user and supposed to be representative of its conceptualization (for instance, this corpus can be the documents written by the end-user on a blog or a wiki) ;
- a **weighting of the properties** associated to each concept. The weights quantify the importance of the end-user associates to the properties in the definition of the concept. These weightings are fixed by the end-user as follows: for each property $p \in P$, the user ordinates, on a 0 to 1 scale, all the concepts having p, in order to reflect its perception on how $p$ is important for defining $c$. For instance, for the property has an author, the concept *Scientific Article* will be put first, secondly the concept *Newspaper Article*, for which the author is less important, thirdly the concept *Technical Manual*.

Thus, SEMIOSEM corresponds to an aggregation of three components:

- an *intentional* component based on the comparison of the properties of the concepts;
- an *extensional* component based on the comparison of the instances of the concepts;
- an *expressional* component based on the comparison of the terms used to denote the concepts and their instances.

Each component of SEMIOSEM is weighted depending on the way the end-user apprehends the domain he/she is working on (e.g. giving more importance to the intentional component when the end-user better apprehends the domain via an intentional approach rather than an extensional one). These differences of importance are conditioned by the domain, the cognitive universe of the end-user and the context of use (*e.g.* ontology-based information retrieval). For instance, in the domain of the animal species, a zoologist will tend to conceptualize them in intention (via the biological properties), whereas the majority of people use more extensional conceptualizations (based on the animals they have met during their life). Formally, the function SEMIOSEM: $C \times C \rightarrow [0, 1]$ is defined as follows:

$$SemioSem(c_1, c_2) = [\alpha * intens(c_1, c_2) + \beta * extens(c_1, c_2) + \gamma * express(c_1, c_2)]^{\frac{1}{\delta}} \quad (22)$$

The following sections present respectively the functions *intens* (cf. Section 4–C1), *extens* (cf. Section 4–C2), *express* (cf. Section 4–C3) and the parameters α, β, γ, δ (cf. Section 4–C4).

1) *Intentional component*: From an intentional point of view, our work is inspired by [12] and is based on the representation of the concepts by vectors in the space of the properties. Formally, to each concept $c \in C$ is associated a vector $!vc = (v_{c1}, v_{c2}, \ldots, v_{cn})$ where $n = |P|$ and $v_{ci} \in [0; 1]$; $\forall i \in [1; n]$. $v_{ci}$ is the weighting fixed by the end-user which precises how the property $i$ is important for defining the concept $c$ (by default, $v_{ci}$ is equal to 1). Thus, the set of concepts corresponds to a point cloud defined in a space with $|P|$ dimensions. We calculate a prototype vector of $c_p$, which was originally introduced in [12] as the average of the vectors of the subconcepts of $c_p$. However [12] only considers the direct subconcepts of $c_p$, whereas we extend the calculation to all the sub-concepts of $c_p$, from generation 1 to generation $n$. Indeed, some properties which can only be associated to indirect subconcepts can however appear in the prototype of the superconcept, in particular if the intentional aspect is important. Thus, the prototype vector *pcp* is a vector in the space properties, where the importance of the property $i$ is the average of the importances of the properties of all the sub-concepts of $c_p$ having $i$. If for $i \in P$, $S_i(c) = \{c_j \leq^C c, c_j \in dom(i)\}$, then:

$$\vec{p}_{cp}[i] = \frac{\sum_{cj \in Si(cp)} \vec{v}_{cj}[i]}{\left|S_i(c_p)\right|} \quad (23)$$

From an intentional point of view, the more the respective prototype vectors of $c_1$ and $c_2$ are close in terms of euclidean distance (*i.e.* the more their properties are close), the more $c_1$ and $c_2$ are similar. This evaluation is performed by the function *intens*: $C \times C \rightarrow [0, 1]$, which is formally defined as follows:

$$intens(c_1, c_2) = 1 - dist(\vec{p}_{c1}, \vec{p}_{c2}) \quad (24)$$

2) *Extensional component*: From an extensional point of view, our work is based on the Jaccard's similarity (cf. Section 4–B). Formally, the function extens: $C \times C \rightarrow [0, 1]$

 

is defined as follows:

$$extens(c_1, c_2) = \frac{|\sigma(c_1) \cap \sigma(c_2)|}{|\sigma(c_1)| + |\sigma(c_2)| - (|\sigma(c_1) \cap \sigma(c_2)|)} \quad (25)$$

This function is defined by the ratio between the number of common instances and the total number of instances minus the number of having common instances. Thus, two concepts are similar when they have a lot of instances in common and few distinct instances.

3) *Expressional component:* From an expressional point of view, the more the terms used to denote the concepts $c_1$ and $c_2$ are present together in the same documents of the corpus, the more $c_1$ and $c_2$ are similar. This evaluation is carried out by the function *express*: $C \times C \rightarrow [0, 1]$ which is formally defined as follows:

$$express(c_1, c_2) = \sum_{t_1, t_2} \left( \frac{min(count(t_1), count(t_2))}{N_{occ}} * \frac{count(t_1, t_2)}{N_{doc}} \right) \quad (26)$$

With :

- $t_1 \in words(c_1)$ and $t_2 \in words(c_2)$ where $words(c)$ returns all the terms denoting the concept c or one of its sub-concept (direct or not);
- $count(t_i)$ returns the number of occurrences of the term ti in the documents of the corpus ;
- $count(t_1, t_2)$ returns the number of documents of the corpus where the term $t_1$ and $t_2$ appear simultaneously;
- $N_{doc}$ returns the number of documents of the corpus;
- $N_{occ}$ is the sum of the numbers of occurrences of all the terms included in the corpus.

4) *Parameters of SEMIOSEM*: $\alpha$, $\beta$, $\gamma$ are the (positive or null) weighting coefficients associated to the three components of SEMIOSEM. In a way of standardization, we impose that the components vary between 0 and 1 and that $\alpha+\beta+\gamma=1$. The values of these three coefficients can be fixed arbitrarily, or evaluated by experiments. We also advocate a method to automatically calculate approximations of these values. This method is based on the following principle. As shown by Figure 1, we consider that the relationship between $\alpha$, $\beta$, $\gamma$, characterizes the cognitive coordinates of the end-user in the semiotic triangle. To fix the values of $\alpha$, $\beta$, $\gamma$, we propose to systematically calculate $\gamma/\alpha$, and $\gamma/\beta$, and then to deduce *alpha* from the constraint $\alpha+\beta+\gamma=1$. $\gamma/\alpha$ (resp. $\gamma/\beta$) is approximated by the cover rate of the concepts (resp. the instances) within the corpus. This rate is equal to the number of concepts (resp. instances) for which at least one of the terms appears in the corpus divided by the total number of concepts (resp. instances). The factor $\delta \geq 0$ aims at taking the mental state of the end-user into account. Multiple works have been done in Cognitive Psychology on the relationship between human emotions and judgments [8]. The conclusion of these works can be summarized as follows: when we are in a negative mental us on what appears to be the more important from an

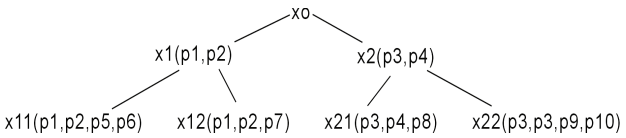


**Figure 1. Intuitive example.**

state (e.g. fear or nervous breakdown), we tend to focus emotional point of view. Respectively, in a positive mental state (e.g. love or joy), we are more open-minded in our judgment and we accept more easily the elements which are not yet be considered as so characteristic. According to [10], a negative mental state supports the reduction in the value of representation, and conversely for a positive mental state. In the context of our measure, this phenomenon is modelised as follows. We characterize 1) a *negative* mental state by a value $\delta \in ]1,+\infty[$, 2) a *positive* mental state by a value $\delta \in ]0,1[$, and 3) a neutral mental state by a value of 1. Thus, a low value of $\delta$, which characterizes a positive mental state, leads to increase the similarity values of concepts which initially would not been considered as so similar. Conversely, a strong value of $\delta$, which characterizes a negative mental state, leads to decrease these values.

## 5. Experimental Results

### 5.1. Comparisons with Human Judgment

In order to evaluate the relevance of our gradients, we have done a first experiment with a group of 19 students of the University of Nantes. The experiment mainly consisted in 1) building by hand a vernacular domain ontology from texts, 2) personalizing this ontology for each student by using the prototypicality gradients, and 3) comparing the results which have been computed to the judgment of each student (*i.e.* its *personal categorization*). The first step of the process was focused on the construction of a vernacular domain ontology. The do main which has been adopted is delimited by the *Grenelle Environment Round Table*, an open multi-party debate in France that brings together representatives of national and local governments and organizations. The aim of this roundtable is to define the key points of a public policy on ecological and sustainable development issues for the next 5 years. Each student has selected 15 texts from several web sites (e.g. industry, political party, professional associations or non-governmental organizations).

Then, for each text, each student has stated a list of the most salient terms (between 10 and 15 per text). The union of all the terms selected individually has conducted to a set of 350 terms[6] which clearly denote the

[6]These terms are only considered as relevant in the following context: 1) a specific endogroup (composed of our 19 students), 2) a delimited domain and 3) a corpus (composed of the texts selected by the students).

 

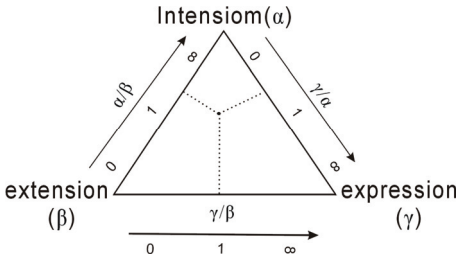


**Figure 2. The weighting coefficients considered as the cognitive coordinates of the end-user in the semiotic triangle. $\gamma/\alpha$ near to 0 indicates that the end-user apprehends the domain in intention (and not in expression); the same ratio close to 1, indicates the opposite and a ratio close to 1 indicates that the intentional and expressional approaches are equilibrated. A similar interpration is adopted for the other ratios. When the three approaches are equilibrated, $\alpha=\beta=\gamma=1/3$ and the cognitive coordinates of the end-user correspond to the barycenter of the semiotic triangle.**

concepts included in the *Grenelle Environment Round Table*. From this terms, we have build a hierarchy composed of 130 concepts (depth = 3, max width = 9). Figure 2 presents an extract of this hierarchy. The second step the process was focused on the personalization of this ontology for each student via the calculation of the prototypicality gradients. For this purpose, we have only used the expressional component of our approach (*i.e.* we didn't define properties, nor instances). First, we asked to each student to consider a concept (e.g. the one which owns the most sub-concepts), and then to sort all its subconcepts by order of typicality. Then, each student has selected several texts (extracted from the web) which, for him, clearly concern the *Grenelle Environment Round Table* and clearly respect its opinions and convictions on this subject. From these texts, each student has calculated the expressional component between all the sub-concepts and the considered concept. The last step of the process was focused on the comparison of the results ($spg_{G;D}$ value) to human judgment (cf. Table 1). (89%) (17 students) have obtained results completely similar or very close to their opinion, and 11% (2 students) obtained different results because their personal corpus didn't really match with their real vision of the subject. This experiment reveals us that 1) the quality of the personalization process is dependent on the composition of the textual corpus: can we consider all the documents selected by the end-user such as mails, texts, web sites or blogs? and 2) the personalization process necessarily requires an adapted corpus. We currently deal with another experiment in collaboration with the University of Bretagne Sud. The main objective of this experiment is to measure the influence of the emotions on cognitive processes, such as for instance how arousal influences the perception of the less typical concepts.

[7]Lucene is a high-performance, full-featured text search engine library written entirely in Java. Lucene is an open source project available at http://lucene.apache.org/.

**Table 1. Values of spg for a student, with the concept *Resource*.**

| c | $spg_{GD}$(resource,c) | Human judgement |
|---|---|---|
| *energy* | 0.93 | 1st |
| *water* | 0.41 | 2nd |
| *money* | 0.01 | 3rd |
| *material* | 0.03 | 4th |

## 5.2. TOOPRAG: A Tool Dedicated to the Pragmatics of Ontology

TOOPRAG (A Tool dedicated to the Pragmatics of Ontology) is a tool dedicated to the automatic calculation of our gradients. This tool, implemented in Java 1.5, is based on Lucene[7] and Jena[8]. It takes as inputs 1) an ontology represented in OWL 1.0, where each concept is associated to a set of terms defined via the primitive rdfs:label (for instance, *<rdfs:label xml:lang="EN"> farmer </rdfs:label>*) and 2) a corpus composed of text files. Thanks to the Lucene API, the corpus is first indexed. Then, the ontology is loaded in memory (via the Jena API) and the SPG values of all the is-a links of the concepts hierarchies are computed. The LPG values of all the terms used to denote the concepts are also computed.

These results are stored in a new OWL file which extends the current specification of OWL 1.0. Indeed, as shown by Figure 4, a LPG value is represented by a new attribute *xml:lpg* which is directly associated to the primitive *rdfs: label*.For instance, the LPG values of the terms "*grower*" and "*peasant*", used to denote the concept "*agricultural labour force*" (*<owl:Class rdf:ID= "agricultural-labour-force">*), are respectively 0.375 and 0. In a similar way, a SPG is represented by a new attribute *xml:cpg*[9] which is directly associated to the primitive *rdfs:subClassOf*. For instance, the SPG values of the is a links defined between the superconcept "*working-population-engaged-in-agriculture*" and its subconcepts "*agricultural-labour-force*", "*farmer*", "*forestranger*" and "*agricultural-adviser*" are respectively 0.0074, 0.9841, 0 and 0.

## 5.3. Distributional Analysis of the SPG

In order to analyze the statistical distribution of the SPG values on different types of hierarchies of concepts,

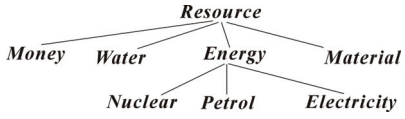


**Figure 3. Extract of a hierarchy of concepts about grenelle environment round table.**

[8]Jena is a Java framework for building Semantic Web applications. It provides a programmatic environment for RDF, RDFS, OWL and SPARQL and includes a rule-based inference engine. Jena is an open source project available at http://jena.sourceforge.net/.

[9]This attribute is called CPG for Conceptual Prototypicality Gradient; SPG and CPG are synonyms.

 

doi:10.4236/iim.2010.22009 Published Online February 2010 (http://www.scirp.org/journal/iim)

```
…
<owl:Class rdf:ID="agricultural labour force">
  <rdfs:label xml:lang="EN" xml:lpg=0.7>farm worker</rdfs:label>
  <rdfs:label xml:lang="EN" xml:lpg=0.3>agricultural labour force</rdfs:label>
  <rdfs:subClassOf rdf:resource="#working population engaged in agriculture" xml:cpg=0.0074/>
</owl:Class>

<owl:Class rdf:ID="farmer">
  <rdfs:label xml:lang="EN" xml:lpg=0.375>grower</rdfs:label>
  <rdfs:label xml:lang="EN" xml:lpg=0.0>peasant</rdfs:label>
  <rdfs:label xml:lang="EN" xml:lpg=0.0>raiser</rdfs:label>
  <rdfs:label xml:lang="EN" xml:lpg=0.625>farmer</rdfs:label>
  <rdfs:subClassOf rdf:resource="#working population engaged in agriculture" xml:cpg=0.9841/>
</owl:Class>

<owl:Class rdf:ID="forest ranger">
  <rdfs:label xml:lang="EN" xml:lpg=0.0>forest ranger</rdfs:label>
  <rdfs:subClassOf rdf:resource="#working population engaged in agriculture" xml:cpg=0.0/>
</owl:Class>

<owl:Class rdf:ID="agricultural adviser">
  <rdfs:label xml:lang="EN" xml:lpg=0.0>agricultural adviser</rdfs:label>
  <rdfs:subClassOf rdf:resource="#working population engaged in agriculture" xml:cpg=0.0/>
</owl:Class>
…
```

**Figure 4. Extract of an OWL file produced by TOOPRAG.**

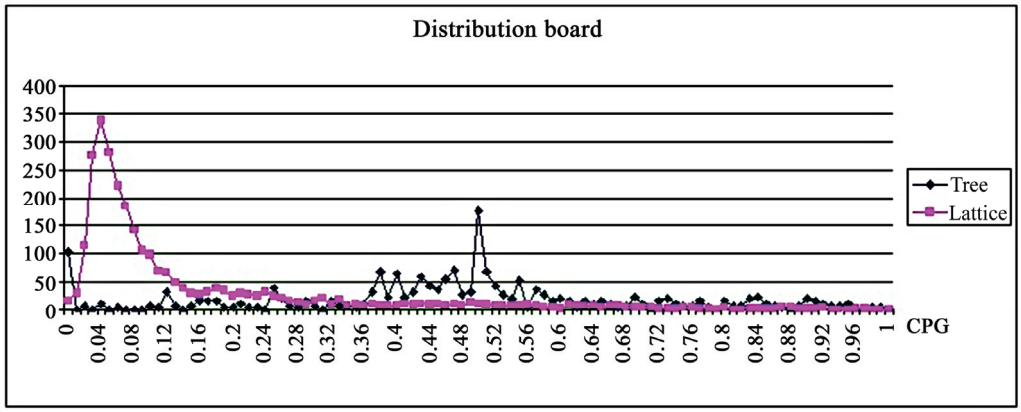


**Figure 5. Influence of the number of edges (with a constant number of concepts).**

we have developed a specific simulator whose parameters (given an ontology $O$) are: $N$ the number of concepts of $O$, H the depth of $O$, and W the max width of $O$. From these parameters, the prototype automatically generates a random hierarchy of concepts. The results presented in Figure 5 computed in the following context: 1) a hierarchy $O_1$ based on a tree described by (N=800, H=9, W= 100), 2) a hierarchy $O_2$ based on a lattice with a density of 0.5 described by (N=800, H=9, W= 100); and (3) $\alpha$=0.3, $\beta$=0.3, $\gamma$=0.3, $\delta$=1. The results clearly attest the fact that multiple inheritance leads to a dilution of the typicality notion.

The results presented in Figure 5 have been calculated in the following context: 1) a hierarchy $O_1$ based on a tree described by (N=800, H=9, W= 100); 2) a hierarchy $O_2$ based on a tree described by (N=50, H=2, W= 30); and 3) $\alpha$=0.3, $\beta$=0.3, $\gamma$=0.3, $\delta$=1, of the distribution of SPG values, proportionally to the size of the hierarchies, for a same density of graphs.

The results presented in Figure 6 have been calculated in the following context: 1) a hierarchy $O$ based on a lattice with a density of 0.66 described by (N=13000, H=7, W= 240), and 2) $\alpha$=0.3, $\beta$=0.3, $\gamma$=0.3, $\delta \in [0,1]$. These results clearly show the relevance of our emotional parameter: in a negative mental state, the distributional analysis focuses on strong values of SPG and in a positive mental state, the distributional analysis is more uniform.

## 5.4. Application of Our SPG: A Case Study in Agriculture

TOOPRAG has been used in a project dedicated to the analysis of texts describing the Common Agricultural

 

Policy (CAP) of the European Union. In this project, we have defined a specific ontology from the multilingual

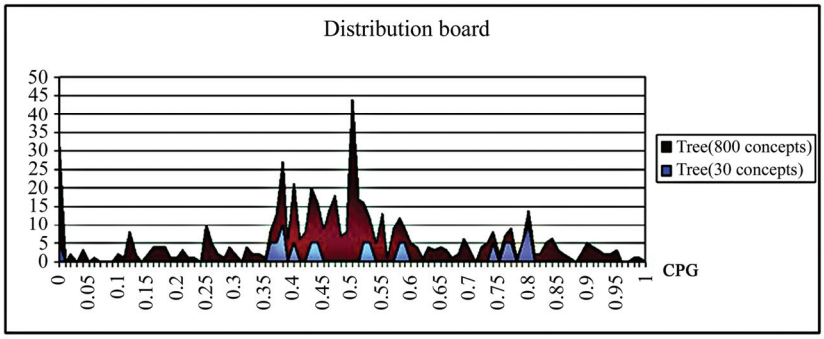


**Figure 6. Influence of the number of concepts in a tree.**

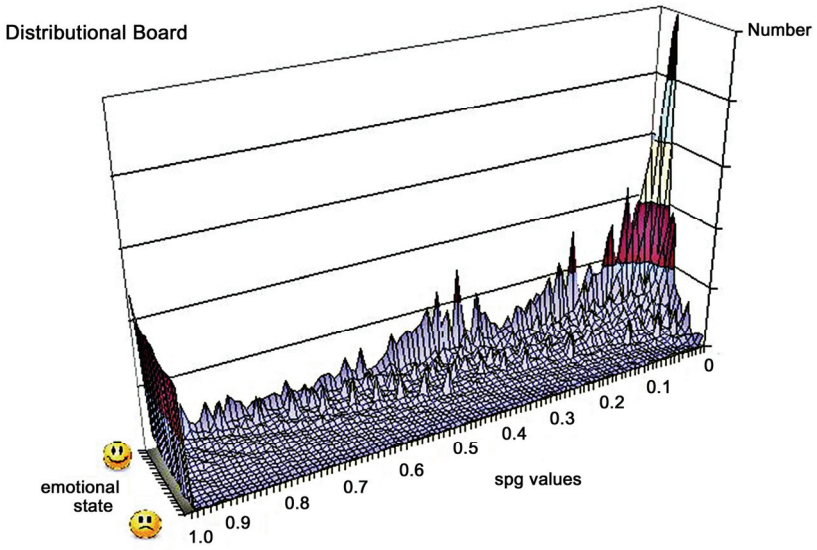


**Figure 7. Emotional parameter influence.**

thesaurus Eurovoc (http://europa.eu/eurovoc/). This thesaurus, which exists in 21 official languages of the European Union, covers multiple fields (e.g. politics, education and communications, science, environment, agriculture, forestry and fisheries, energy, etc.). It provides a means of indexing the documents in the documentation systems of the European institutions and of their users (e.g. the European Parliament, some national government departments and European organizations). From the Eurovoc field dedicated Agriculture, we have defined a first hierarchy of concepts by using the hyponymy/hyperonymy relationships (identified by the "Broader Term" links in Eurovoc) and the synonymy relationships (identified by the "Used For" links in Eurovoc). Then, this hierarchy has been modified and validated by an expert in Agriculture and Forestry. In its current version, this ontology includes a hierarchy of concepts based on a tree described by 283 concepts (depth=4 and max width=11). The lexicon of this ontology is composed of 597 terms. In average, each concept is associated to 2,1 terms (min=1 and max=11). The corpus used for this experimentation is composed of 55 texts published in the Official Journal of the European Union (http://eur-lex. europa.eu) since 2005, and in particular 43 regulations, 1 directive, 8 decrees, 3 community opinions. It includes 1.360.000 words. From a statistical point of view, 61 concepts of the ontology are directly evocated in the corpus through the terms and 37 indirectly (via inheritage). Thus, the reverse ratio is 34,63%. Although the ontology considered in this project does not yet include properties and instances, the results provided by TOOPRAG (in the context of this specific corpus) are interesting because they help the expert to analyze the Common Agricultural Policy (CAP) of the European Union through regulatory texts. For instance, as shown by the Figure 8, the SPG values clearly underline that since 2005, the cultivation system which is particularly encouraged by the PAC is the organic farming.

 

In a similar way, the SPG values presented in Figure 4 state that the blue-collar workers of the agricultural sector (e.g. farm workers, farmers or peasants) are more supported by the PAC than the white-collar workers (e.g. agricultural advisers or head of agricultural holdings).

## 5.5. Application of our SPG: A Case Study in "Health, Safety and Environment"

Our approach is currently evaluated in the context of a project[10] dedicated to Legal Intelligence within regulatory documents related to the domain "Health, Safety and Environment" (HSE). A first ontology of this domain has been defined[11]. In its current version, it is composed of 3776 concepts structured in a lattice-based hierarchy (depth = 11 ; width = 1300). The calculation of our gradient has been applied on a specific corpus which includes 2052 texts related to the HSE domain from a reglementary point of view[12]: 782 administrative orders, 347 circulars, 288 parts of regulation, 178 decrees, 139 parts of legal rules, 102 european directives, 32 laws, 11 memorandum, 4 ordinances, etc. This process indicates that 1) 30.2% of the SPG values are non-null, 2) 3.34% of the SPG values are equal to 1, 3) 6.18% of the SPG values belong to [0.5, 1] and 4) 63.23% of SPG values belong to [0, 0.01]. The median value of the GPS is equal to 0.128.

## 5.6. Application of SEMIOSEM

SEMIOSEM is currently evaluated with the ontology of the HSE domain. In order to evaluate our measure and to compare it with related work, we have focused our study on the hierarchy presented in Figure 9: the goal is to compute the similarity between the concept *Carbon* and the sub-concepts of *Halogen*; for the expert of Tennaxia, these similarities are evaluated as follows: *Fluorine*=0.6; *Chlorine*=0.6; *Bromine*=0.3; *Iodine*=0.3; *Astatine*=0.1. We have also elaborated a specific corpus of texts composed of 1200 european regulatory documents related to the HSE domain (mainly laws, regulations, decrees and directives).

Table 2 presents the similarity values obtained with three intentional-based measures: Rada, Leacock and Wu. One can note that all the values are equal because these measures only depend on the structure of the hierarchy.

Table 3 depicts the similarity values obtained with three extensional-based measures: Lin, Jiang and Resnik. Table 4 presents the similarity values obtained with SEMIOSEM according to 6 contexts defined by the following parameters:

```
…
<owl:Class rdf:ID="organic farming">
  <rdfs:label xml:lang="EN" xml:lpg=1.0>organic farming</rdfs:label>
  <rdfs:subClassOf rdf:resource="#cultivation system" xml:cpg=0.7/>
</owl:Class>

<owl:Class rdf:ID="intensive farming">
  <rdfs:label xml:lang="EN" xml:lpg=0.0>intensive farming</rdfs:label>
  <rdfs:subClassOf rdf:resource="#cultivation system" xml:cpg=0.0/>
</owl:Class>

<owl:Class rdf:ID="single-crop farming">
  <rdfs:label xml:lang="EN" xml:lpg=0.0>single-crop farming</rdfs:label>
  <rdfs:subClassOf rdf:resource="#cultivation system" xml:cpg=0.0/>
</owl:Class>

<owl:Class rdf:ID="extensive farming">
  <rdfs:label xml:lang="EN" xml:lpg=1.0>extensive farming</rdfs:label>
  <rdfs:subClassOf rdf:resource="#cultivation system" xml:cpg=0.0333/>
</owl:Class>

<owl:Class rdf:ID="dry farming">
  <rdfs:label xml:lang="EN" xml:lpg=0.0>dry farming</rdfs:label>
  <rdfs:subClassOf rdf:resource="#cultivation system" xml:cpg=0.0/>
</owl:Class>

<owl:Class rdf:ID="crop rotation">
  <rdfs:label xml:lang="EN" xml:lpg=1.0>crop rotation</rdfs:label>
  <rdfs:subClassOf rdf:resource="#cultivation system" xml:cpg=0.2667/>
</owl:Class>
…
```

**Figure 8. Extract of an OWL file produced by TOOPRAG.**

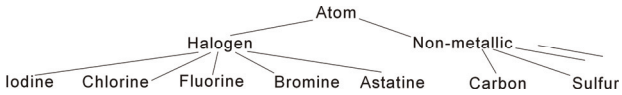


**Figure 9. Extract of the hierarchy of concepts of the HSE ontology.**

- A ($\alpha$=0.7, $\beta$=0.2, $\gamma$=0.1, $\delta$=1);
- B ($\alpha$=0.2, $\beta$=0.7, $\gamma$=0.1, $\delta$=1);
- C ($\alpha$=0.2, $\beta$=0.1, $\gamma$=0.7, $\delta$=1);
- D ($\alpha$=0..33, $\beta$=0.33, $\gamma$=0.33, $\delta$=0.1);
- E ($\alpha$=0.7, $\beta$=0.2, $\gamma$=0.1, $\delta$=0.1);
- F ($\alpha$=0.7, $\beta$=0.2, $\gamma$=0.1, $\delta$=5.0);

These experimental results lead to the following remarks:

- in all the contexts, SEMIOSEM provides the same order of similarities as the other measures. In a context which gives priority to the intentional component (cf. Context A), SEMIOSEM is better than the other measures. In the context B which gives priority to the extensional component (resp. the context C which gives priority to the expressional component), SEMIOSEM is close to Jiang's measure (resp. Lin's measure). In a context

**Table 2. Similarity with Carbon (Rada, Leacock, Wu).**

| Halogen | *Rada* | *Leacock* | *Wu* |
|---|---|---|---|
| *Fluorine* | 0.25 | 0.097 | 0.6 |
| *Chlorine* | 0.25 | 0.097 | 0.6 |
| *Bromine* | 0.25 | 0.097 | 0.6 |

[10]This ongoing research project is funded by the French company Tennaxia (http://www.tennaxia.com). This "IT Services and Software Engineering" company provides industry-leading software and implementation services dedicated to Legal Intelligence.

[11]INPI June 13, 2008, Number 322.408 – SCAM-Velasquez September 16, 2008, Number 2008090075. All rights reserved.

[12]These texts have been extracted from the European Parliament and the French Parliament, mainly from LegiFrance (http://www.legifrance.gouv.fr/) and Eur-Lex (http://eur-lex.europa.eu/).

 

| | | | |
|---|---|---|---|
| *Iodine* | 0.25 | 0.097 | 0.6 |
| *Astatine* | 0.25 | 0.097 | 0.6 |

**Table 3. Similarity with carbon (Lin, Jiang, Resnik).**

| Halogen | *Lin* | *Jiang* | *Resnik* |
|---|---|---|---|
| *Fluorine* | 0.31 | 0.14 | 1.43 |
| *Chlorine* | 0.28 | 0.12 | 1.43 |
| *Bromine* | 0.23 | 0.09 | 1.43 |
| *Iodine* | 0.22 | 0.09 | 1.43 |
| *Astatine* | 0 | 0 | 1.43 |

**Table 4. Similarity with carbon (SemioSem).**

| Halogen | *A* | *B* | *C* | *D* | *E* | *F* |
|---|---|---|---|---|---|---|
| *Fluorine* | 0.40 | 0.14 | 0.32 | 0.27 | 0.91 | 0.025 |
| *Chlorine* | 0.36 | 0.12 | 0.29 | 0.25 | 0.90 | 0.017 |
| *Bromine* | 0.29 | 0.10 | 0.23 | 0.20 | 0.88 | 0.007 |
| *Iodine* | 0.28 | 0.10 | 0.23 | 0.19 | 0.88 | 0.006 |
| *Astatine* | 0.01 | $2.10^{-4}$ | $2.10^{-4}$ | $3.10^{-4}$ | 0.63 | $1.10^{-8}$ |

which gives no priority to a specific component (cf. Context D), SEMIOSEM is between Lin's measure and Jiang's measure;

- context E and F clearly show the influence of the emotional factor: a positive mental state (cf. context E) clearly increases the similarities values and a negative mental state (cf. context F) clearly decreases similarities values;
- the concept Astatine is not evocated in the corpus, nor represented by instances. Thus, it is not considered as similar by Lin's and Jiang's measures. SEMIOSEM finds a similarity value thanks to the intentional component.

## 6. Discussion and Future Work

### 6.1. Lexical and Conceptual Prototypicality Gradients

The purpose of our work, which is focused on the notion of "Personalized Vernacular Domain Ontology", is to deal with subjectivity knowledge via 1) its specificity to an endogroup and a domain, 2) its ecological aspect and 3) the prominence of its emotional context. This objective leads us to study the pragmatic dimension of an ontology. Inspired by works in Cognitive Psychology, we have defined two measures identifying two complementary gradients called SPG and LPG which are respectively dedicated to 1) the conceptual prototypicality which evaluates the representativeness of a concept within a decomposition and 2) the lexical prototypicality which evaluates the representativeness of a term within a set of terms used to denote a concept. It is important to underline that these gradients do not modify the formal semantics of the ontology which is considered; the subsumption links remain valid. These gradients only reflect the pragmatics of an ontology for knowledge (re)-using. They can be an effective help in different activities, such as:

- *Information Retrieval*. Our conceptual and lexical prototypicality gradients can be used to classify the results of a query, and more particularly an extended query, according to a relevance criteria which consists in considering the most representative element of a given concept (resp. a given term) as being the most relevant result of a query expressed by a (set of) term(s) denoting this concept (resp. corresponding to this term). This approach permits a classification of the extended results from a qualitative point of view. Moreover, our approach also allows us to proportion the number of results according to the value of the gradients (*i.e.* a quantitative point of view). Thus, information retrieval becomes customizable, because it is possible to adapt the results to the pragmatics of the ontology, i.e. privileging the intentional dimension (and not the extensional and expressional one) or conversely, working with different mental states, etc. In this way, *Ontology Personalization* is used as a means for Web Personalization and Semantic Web Personalization. We currently evaluate this approach in the context of the SweetWiki Semantic Web platform [24].
- *Ontological Analysis* of text Corpora. As introduced in Section 5-D, our Semiotic-based Prototypicality Gradient can be used to evaluate the ontological contents of text corpora: which are the main concepts involved in a text corpus? By using the same ontology applied on different corpora related to the same domain, it is possible to compare, at the conceptual level, the Information Content of these corpora. We currently evaluate this approach in the context of an experimentation which aims at making a comparative analysis of the health-care preoccupations of different populations, in particular French, English and American population. For this purpose, we currently define an multilingual ontology from the MeSH[13]. In order to really deal with the preoccupations of people, we have selected the three most popular and complete medical websites where french, english and american people can find and exchange all the information they are looking about their health care needs: Doctissimo in France[14], Health-care Republic in United Kingdom[15] and HealthCare.com in USA[16]. These websites will be considered as three distinctive corpora from which our gradients will be calculated, by using the same multilingual ontology.

Our work is currently in progress towards the improvement of the gradients according to many works related to Cognitive and Social Psychology. We also study how to enrich the intentional component by taking the axiomatic part of an ontology into account. From an application point of view, we currently evaluate our approach in the context of a project dedicated to Legal Intelligence within regulatory documents related to the areas "Hygiene, Safety and Environment".

[13]MeSH is the U.S. National Library of Medicine's controlled vocabulary (http://www.nlm.nih.gov/mesh/meshhome.html). MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts.
[14]http://www.doctissimo.fr

 

[15]http://www.healthcarerepublic.com
[16]http://www.healthcare.com

### 6.2. SemioSem

SEMIOSEM is particularly relevant in a context where the perception (by the end-user) of the domain which is considered (and which is both conceptualized within an ontology and expressed by a corpus and instances) can have a large influence on the evaluation of the similarities between concepts (e.g. ontology-based information retrieval). We advocate that such an user-sensitive context, which de facto includes subjective knowledge (whereas ontologies only includes objective knowledge), must be integrated in a similarity measure since ontologies co-evolve with their communities of use and human interpretation of context in the use. Formally, SEMIOSEM respects the properties of similarity measure reminded in [17]: *positiveness*[17], *reflexivity*[18] and *symmetry* [19]. But, SEMIOSEM is not a semantic distance since it does not check simultaneously the strictness property[20] and the triangular inequality[21]. For the extensional component, our first choice was the Amato measure. But one of our goal is to be independent from the modeling structure and this measure clearly depends on the Most Specific Common Subsumer (MSCS). Moreover, in the case of our experiment, it does not really provide more relevant results. This is why we have selected the Jaccard measure for its simplicity and its independence from the MSCS but we currently study the use of the Dice measure. For the expressional component, the Latent Semantic Analysis could be adopted but since it is based on the tf-idf approach, it is not really appropriated to our approach: we want to keep the granularity of the corpus in order to give more importance to concepts which perhaps appears less frequently in each document, but in an uniform way in the whole corpus (than concepts which are frequently associated in few documents). Then, as we simply compare the terms used to denote the concepts in the corpus, our approach is clearly limited since it can not deal with expressions such as $t_1$ and $t_2$ are opposite. To deal with this problem, we plan to study more sophisticated computational linguistic methods. Finally, for the intentional component, our approach can be timeconsuming (when the end-user decides to weight the properties of the concepts[22]), but, as far as we know, it is really innovative and it provides promising results. The parameters alpha, beta, gamma and delta are used to adapt the measure to the context which is related to the end-user perception of the domain according to the intentional, extensional, expressional and emotional dimension. We consider that this is really a new approach and this is why we call our measure SEMIOSEM (Semiotic-based Similarity Measure). Moreover, the aggregation we advocate does not just correspond to a sum: these parameters are used both to adapt the influence of each dimension and/or to adapt the calculus according to the resources with are available. Thus, when no corpus is available, the expressional component can not be used ($\gamma$=0). A similar approach is adopted for the intentional component (an ontology without properties leads to $\alpha$ = 0) and the extensional component (no instances leads to $\beta$ = 0). The value of delta (emotional state) can be defined according to a questionary or the analysis of data given by physical sensors such as the speed of the mouse, the webcam-based facial recognition, etc. To sum up, SEMIOSEM is more flexible (since it can deal with multiple information sources), more robust (since it performs relevant results under unusual conditions as shown by the case Astatine of the experimental results) and more user-centered (since it is based on the domain perception and the emotional state of the end-user) than all the current methods.

[17]$\forall$ x, y $\in$ C :SEMIOSEM(x,y) $\geq$ 0
[18]$\forall$ x, y $\in$ C :SEMIOSEM(x,y) $\leq$ SEMIOSEM(x,y)
[19]$\forall$ x, y $\in$ C :SEMIOSEM(x,y)=SEMIOSEM(y,x)
[20]$\forall$ x, y $\in$ C :SEMIOSEM(x,y)=0 $\Rightarrow$x = y
[21]$\forall$ x, y,z $\in$ C :SEMIOSEM(x,y)+ SEMIOSEM(y,z) $\geq$ SEMIOSEM(x,z)

## 7. References

[1] S. Harnad, "Categorical perception," Encyclopedia of Cognitive Science, Vol. LXVII, No. 4, 2003. [Online]. Available: http://cogprints.org/3017/.

[2] T. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," in Formal Ontology in Conceptual Analysis and Knowledge Representation, N. Guarino and R. Poli, Eds. Deventer, The Netherlands: Kluwer Academic Publishers, 1993.

[3] C. K. Ogden and L. Richards, "The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism," Harcourt, iSBN–13: 978–0156584463, 1989.

[4] D. L. M. Gabora, D. E. Rosch, and D. D. Aerts, "Toward an ecological theory of concepts," Ecological Psychology, Vol. 20, No. 1–2, pp. 84–116, 2008. [Online]. Available: http://cogprints.org/5957/.

[5] E. Rosch, "Cognitive reference points," Cognitive Psychology, No. 7, pp. 532–547, 1975.

[6] M. McEvoy and D. Nelson, "Category norms and instantce norms for 106 categories of various sizes," American Journal of Psychology, Vol. 95, pp. 462–472, 1982.

[7] C. Morris, "Foundations of the theory of signs," Chicago University Press, 1938.

[8] S. Bluck and K. Li, "Predicting memory completeness and accuracy: Emotion and exposure in repeated autobiographical recall," Applied Cognitive Psychology, No. 15, pp. 145–158, 2001.

[9] J. Park and M. Nanaji, "Mood and heuristics: The influence of happy and sad states on sensitivity and bias in stereotyping," Journal of Personality and Social

[22]Again, by default, all the weightings are equal to 1 and the function Intens remains valid. In the case of our experiment, the results obtained in this context for the concept Fluorine are: A – 0.59; B – 0.19; C – 0.38 D – 0.37; E – 0.95; F – 0.12.

 

Psychology, No. 78, pp. 1005–1023, 2000.

[10] M. Mikulincer, P. Kedem, and D. Paz, "Anxiety and categorization-1, the structure and boundaries of mental categories," Personnality and Individual Differences, Vol. 11, No. 11, pp. 805–814, 1990.

[11] C. M. Au Yeung and H. F. Leung, "Formalizing typicality of objects and context-sensitivity in ontologies," in AAMAS '06: Proceedings of the fifth international joint conference on Autonomous Agents and Multiagent Systems. New York, NY, USA: ACM, ISBN 1-59593-303-4, pp. 946–948, 2006.

[12] C. M. Au Yeung and H. F. Leung, "Ontology with likeliness and typicality of objects in concepts," in Proceedings of the 25th International Conference on Conceptual Modeling–ER 2006, S. B. Heidelberg, Ed., Vol. 4215/2006, ISSN 0302-9743 (Print), 2006.

[13] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in 14th International Joint Conference on Artificial Intelligence (IJCAI 95), Montral, Vol. 1, pp. 448–453, August 1995.

[14] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," Science, No. 185, pp. 1124 –1131, 1974.

[15] D. Lin, "An information-theoric definition of similarity," in Proceedings of the 15th International Conference on Machine Learning, pp. 296–304, 1998.

[16] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxinomy," in International Conference on Research in Computationnal Linguistics, pp. 19–33, 1997.

[17] C. d'Amato, S. Staab, and N. Fanizzi, "On the influence of description logics ontologies on conceptual similarity," in EKAW 2008, International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns, pp. 48–63, October 2008.

[18] R. Rada, H. Mili, E. Bicknell, and M.Blettner, "Development and application of a metric on semantic nets," IEEE Transactions on Systems, Man and Cybernetics, Vol. 19, No. 1, pp. 17–30, 1989.

[19] C. Leacock and M. Chodorow, "WordNet: An electronic lexical database," Cambridge, MA, The MIT Press, 1998, ch. Combining local context and wordnet similarity for word sense identification, pp. 265–283.

[20] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp. 133–138, 1994.

[21] E. Blanchard, M. Harzallah, and P. Kuntz, "A generic framework for comparing semantic similarities on a subsumption hierarchy," in Proceedings of the 18th European Conference on Artificial Intelligence (ECAI'2008). IOS Press, pp. 20–24, 2008.

[22] P. Jaccard, "Distribution de la flore alpine dans le bassin des dranses et dans quelques rgions voisines," Bulletin de la Socit Vaudoise de Sciences Naturelles, Vol. 37, pp. 241–272, 1901, (in French).

[23] M. Sanderson and W. Croft, "Deriving concept hierarchies from text," in Proceedings of the 22nd International ACM SIGIR Conference, pp. 206–213, 1999.

[24] M. Buffa, F. Gandon, G. Ereteo, P. Sander, and C. Faron, "Sweetwiki: A semantic wiki," Special Issue of the Journal of Web Semantics on Semantic Web and Web 2.0, Vol. 6, pp. 84–97, February 2008.

 