On the consideration of a bring-to-mind model for computing the information content of concepts defined into ontologies

Sébastien Harispe, Abdelhak Imoussaten, François Trousset, Jacky Montmain

On the consideration of a *bring-to-mind* model for computing the Information Content of concepts defined into ontologies

Sébastien Harispe, Abdelhak Imoussaten, François Trousset and Jacky Montmain
Laboratory of Computer Science and Production Engineering (LGI2P) Ecole des mines d'Alès,
Parc scientifique G. Besse, 30035 Nîmes cedex 1, France.
Email: {firstname.name}@mines-ales.fr

***Abstract*—Ontologies are core elements of numerous applications that are based on computer-processable expert knowledge. They can be used to estimate the Information Content (IC) of the key concepts of a domain: a central notion on which depend various ontology-driven analyses, e.g. semantic measures. This paper proposes new IC models based on the belief function theoretical framework. These models overcome limitations of existing ICs that do not consider the *inductive inference assumption* intuitively assumed by human operators, i.e. that occurrences of a concept (e.g. *Maths*) not only impact the IC of more general concepts (e.g. *Sciences*), as considered by traditional IC models, but also the one of more specific concepts (e.g. *Algebra*). Interestingly, empirical evaluations show that, in addition to modelling the aforementioned assumption, proposed IC models compete with best state-of-the-art models in several evaluation settings.**

I. Introduction, problem and contributions

Ontologies are central components of a large variety of applications that rely on computer-processable domain expert knowledge, e.g. medical information and clinical decision support systems [1]. In particular, they provide taxonomies defining partial orders of the key concepts of a domain (such as disease classifications). By defining generalization and specialization relationships (i.e. hypernym-hyponym relationships) between concepts, these taxonomies give access to consensual human cognitive views of concept hierarchical relationships. They are therefore of particular interest for designing Artificial Intelligence systems and are largely used in Information Retrieval, Computational Linguistics and Approximate Reasoning to cite a few. An important aspect of taxonomies is that they give the opportunity to analyse intrinsic and contextual properties of concepts. Indeed, by analysing their topologies and additional information about concept usage, several authors have proposed models which take advantage of these taxonomies in order to estimate the *informativeness* or *Information Content* (IC) of concepts [2]. IC models are designed to mimic human, generally consensual and intuitive, appreciation of concept informativeness. As an example, most people will agree that the concept *Algebra* is more informative than the concept *Mathematics* in the sense that knowing the fact *Lucie studies Algebra* is more informative than knowing that *Lucie studies Mathematics* – in this case, this is ensured since the second fact is entailed by the other because *Algebra* is a specific topic of *Mathematics*. Accurate concept informativeness estimators are central for extending ontology usage to applications which do not only depend on exact reasoning. Indeed, various ontology-driven analyses, such as computing the similarity of concepts, extensively depend on accurate IC computational models. For instance, these models play an important role in the definition of semantic measures widely used in Information Retrieval, Knowledge Inference, and Natural Language Processing.

The global aim of defining and using an ontology could be to characterize the topics of interest of users by analysing their book libraries. In this case, analyses will be based on a taxonomy that organizes various book topics into a poset, as specified in Figure 1.

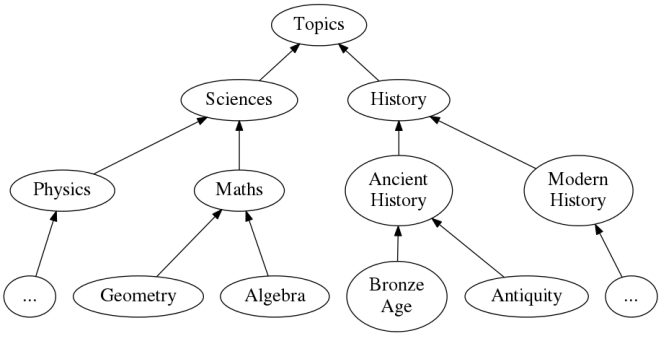

Fig. 1. Taxonomy of book topics.

In this setting, the IC of concepts could be used to characterize the center of interest of a specific user. Knowing that a user has a textbook about *Maths* will be considered less informative than knowing that he has a textbook about *Algebra*. Indeed, due to the transitivity of the taxonomic predicate, the taxonomy specifies that any book which is associated to the topic *Algebra* is also associated to the topics *Maths* and *Sciences*. In other words, when the library of a specific user will be analysed, any books talking about *Algebra* will contribute to increase the consideration that the user has interest for the topics *Algebra*, *Maths* and *Sciences*. In existing IC computational models, the informativeness of the concept *Maths* is only regarded as a function of the number of its (i) direct and (ii) indirect instances, i.e. the cardinality of the set composed of (i) the books explicitly annotated by *Maths* and (ii) the books annotated by *Algebra* and *Geometry* in this case. Therefore, when the IC of a concept is computed using existing models, the information relative to the number of occurrences of its subsumers will not be taken into account, e.g. in Figure 1 the number of occurrences of the concept

Maths will not be taken into account for estimating the IC of the concept *Algebra*. However, this modelling choice can lead to undesired results and may therefore not be adapted to several usage contexts. Let us highlight the core of the problem using a simple example. Considering an extreme case in which a user has a library of 100 books among which: 98 are explicitly annotated by the topic *Maths*, 1 is annotated by *Algebra* and 1 is annotated by *Antiquity*, we will obtain similar IC values for the two topics *Algebra* and *Antiquity*. Otherwise stated, this means that, in order to characterize a user profile, knowing that the user is interested in books talking about *Algebra* is just as much informative as knowing that the user is interested in books talking about *Antiquity*; this, even if we already knew that the user was extensively interested about the topic *Maths* (98 books). In our opinion, this highlights a limit of existing IC computational models which only consider partial information about concepts relationships and observations – in contradiction to what intuition seems to do. Indeed, in such a case, most people will agree that the topic *Antiquity* carries more information than *Algebra* since it enables identifying a potential center of interest of the user that was not *suspected* considering other observations. This paper proposes to define and study new IC models considering the assumption that appraisal of concept informativeness is contextual and highly impacted by the fact that we, cognitively speaking, extensively consider *inductive inference* in daily life. For instance, telling you that someone likes books of *Maths* will tend to reinforce you to think that he may also like books of *Algebra*. For convenience, throughout this paper, we will denote this assumption the *inductive inference assumption*.

This paper proposes and evaluates new IC models that overcome the limitations of existing models underlined above by implementing the *inductive inference assumption*. These models can be used to estimate new ICs which consider occurrences of more general concepts when computing the IC of a concept. This is done by integrating a *bring-to-mind* model into the IC model, by considering that a concept occurrence (*Maths*) may impact the IC of its descendants (*Algebra*). To this end we propose to use well-known contributions made in the belief function theoretical framework.

The paper is structured as follow. Section II introduces the formalism and existing works on which is based our contributions – among others, it presents existing IC computational models. Section III introduces interesting notions of the belief function theoretical framework, and presents how we propose to use them for defining new IC computational models implementing the *inductive inference assumption*. Section IV is dedicated to the evaluation and compares proposed models to existing works. Evaluation is mainly discussed with regard to the impact of the new IC proposals on semantic measures accuracy. Section V summarizes our contributions and concludes this paper.

II. Formalism and existing works

A. Ontologies: Terminology and formalism

We consider an ontology from which can be derived, if necessary after applying inference procedures, a taxonomy $O = (\preceq, C)$ partially ordering ($\preceq$) the concepts it defines (C) – please refer to Figure 1 for an example. We denote $A(c) = \{x \in C | c \preceq x\}$ and $D(c) = \{x \in C | x \preceq c\}$ respectively the inclusive ancestors and inclusive descendants of the concept $c \in C$. The *root* is the unique concept without ancestors ($A(root) = \{root\}$) and a concept without descendant is denoted a leaf ($D(leaf) = \{leaf\}$) ; $leaves \subseteq C$ is the set of leaves. We also denote $leaves_c$ the set of leaves that are subsumed by the concept c, i.e. $leaves_c = D(c) \cap leaves$.

A concept can be considered as a class composed of set of instances, e.g. in Figure 1 the concept *Maths* can be used to refer to the set of books which are annotated by the concept *Maths*. We denote I the set of instances of our domain (books), and $I^*(c) \subseteq I$ the instances that are explicitly annotated by the concept c (without considering any inference procedure based on the concept partial ordering defined by the taxonomy). We consider that no annotation associated to an instance can be inferred, i.e. $\forall (x, y) \in C \times C$, with $x \preceq y$, $I^*(x) \cap I^*(y) = \emptyset$. We denote $I(c) \subseteq I$ the instances that are associated to the concept c considering the transitivity of the taxonomic relationship and concept partial ordering $\preceq$, e.g. $I(Algebra) \subseteq I(Maths)$. We therefore obtain $\forall c \in C, I(c) = \bigcup_{x \in D(c)} I^*(x)$. From these points we can stress that $I(root) = I$, and that $\forall c \in leaves$, $I(c) = I^*(c)$.

B. Existing Information Content models

Due to the transitivity of the taxonomic relationship the instances of a concept $x \in C$ are also instances of any concept subsuming x, i.e. $x \preceq y \Rightarrow I(x) \subseteq I(y)$. This central notion is generally used to discuss the specificity of a concept, i.e. how restrictive a concept is with regard to I. The more restrictive a concept, the more specific it is considered to be. In the literature, the specificity of a concept is also regarded as the Information Content (IC) that is conveyed by a concept; both notions are synonyms. In this paper we will refer to the notion of IC defined through a function $IC : C \rightarrow \mathbb{R}_+$. In accordance to knowledge modelling constraints, any IC function must monotonically decrease from the leaves to the root of the ontology such as $x \preceq y \Rightarrow IC(x) \geq IC(y)$. Two main approaches have been proposed to estimate the IC of a concept, they are presented hereafter.

The *Intrinsic approach* estimates the IC of a concept by making a topological analysis of the taxonomy and by studying the location of its corresponding node in the taxonomy. Among the earlier estimators that have been taken into account, researchers have proposed to estimate IC of a concept based on its depth, its number of ancestors/descendants or the number of leaves it subsumes. As an example, Seco et al. [3] defined the IC of a concept as inversely proportional to its number of descendants:

$$IC_{Seco}(c) = 1 - \frac{log\ |D(c)|}{log\ |C|} \tag{1}$$

More refined expressions have been proposed. As an example, Sanchez et al. [4] propose an approach that considers both the number of leaves a concept subsumes, as well as its number of ancestors. The rationale is that the less the number of leaves a concept subsumes, the more informative it will be. In addition, considering two concepts that subsume the same number of leaves, the concept with the highest number of ancestors will probably be more informative. With $leaves_c$ the leaves that are

subsumed by the concept c, the authors proposed the following expression:

$$IC_{Sanchez}(c) = -log\frac{\frac{|leaves_c \setminus \{c\}|}{|A(c)|} + 1}{|leaves| + 1} \quad (2)$$

Intrinsic ICs are efficient to estimate the informativeness of concepts by analysing the topological properties of their taxonomical ordering. Nevertheless, they cannot be used to take into account concept usage in specific application contexts. However, in several cases, concept informativeness can only be estimated with regard to a specific application context, and must therefore be estimated considering concept usage in this context. As an example, the concept *Fuzzy logic*, unknown by most people, will be considered very specific/informative in most context. However, it will not be considered very informative to characterize articles published at *FuzzIEEE* since most of them will be related to this topic. To overcome this limitation of intrinsic approaches, extrinsic evidence (i.e. that can be found outside the taxonomy) have been taken into account to estimate concept informativeness.

The *Extrinsic approach* is based on Shannon's Information Theory and proposes to assess the informativeness of a concept by analysing a corpus of texts. Originally defined by Resnik [2], the IC of a concept c is defined to be inversely proportional to $p(c)$, the probability that c occurs in a corpus. Considering that evidence of concept usage can be obtained by studying a collection of entities (books) annotated by concepts, the probability that an instance of I belongs to $I(c)$ can be defined such as $p : C \rightarrow [0, 1]$ with $p(c) = |I(c)|/|I|$. The informativeness of a concept is next assessed by defining:

$$IC(c) = -log\ p(c) \quad (3)$$

We have introduced intrinsic and extrinsic approaches that are used to estimate the IC of concepts defined into a taxonomy – additional examples of IC expressions can be found in [5]. Intrinsic formulations cannot take into account context-specific specificities about concept usage. In addition, extrinsic models that overcome this limit only consider information related to the descendants of a concept to estimate its informativeness. Therefore, these models do not implement the *inductive inference assumption* introduced in Section 1: even if a large number of observations have been made for a concept, this will not reinforce our confidence to observe one of its descendants, e.g. *Knowing that Lucie has 98 Math books in her library*, the two additional facts (i) *Lucie has a book on Algebra* and (ii) *Lucie has a book on Antiquity* are both equally informative. Thus, both intrinsic and extrinsic IC models, by design, fail to implement the *inductive inference assumption* we would like to integrate for estimating concept informativeness. The next section introduces our proposal to overcome this limitation.

III. IC Models based on belief function theoretical framework

Let consider the example provided in Section I: knowing that *Lucie studies Algebra* is more informative than knowing that *Lucie studies Maths* – according to the taxonomy of Figure 1 that shows that *Algebra* is a leaf while *Maths* is one of its parent. The statement *Lucie studies Maths* means that the question: *What does Lucie study?* cannot be answered more precisely by the informer. However, we can assume that this statement may only convey an imprecise information about Lucie because of informer lack of knowledge. And that, a more precise answer which could only be provided by an informer with deep knowledge about Lucie, would make clear that Lucie studies *Geometry* or *Algebra*. This is the assumption we will use to implement the *inductive inference assumption* using the belief function theoretical framework.

Prior to introduce our modelling strategy, we stress the two main implications of considering this assumption. First, by considering that an annotation or answer (e.g. *Maths*) is imprecise if it does not refer to a leaf, we assume (A) that all annotations/answers could, in the absolute, always be reduced to a leaf or a set of leaves. This is not always true. As an example, a book talking about "*Ancient Babylonian mathematics*" could refer to the topic mathematics without explicitly discussing (a) specific discipline(s) – saying that annotating this book by the concept *Maths* is imprecise would be inappropriate. In addition, by considering (B) that an imprecise annotation/answer refers to a subset of leaves, we make a *closed-world assumption*, i.e. we consider that all precise answers can be expressed using a subset of leaves of the taxonomy, and that all possible answers are therefore made explicit into the taxonomy. Otherwise stated, it is considered that all answers are covered by the ontology – which therefore is assumed to completely model the domain of interest. This is in contradiction with the *open-world assumption* classically considered in Knowledge Representation and will therefore not be adapted to all cases, e.g. a book talking about *Logics* (a concept not specified into the taxonomy) could not be annotated by *Geometry* or *Algebra*. We have underlined the implications of considering that annotations that are not leaves are imprecise for our modelling strategy. However, we defend that these two implications are intrinsically tied to the *inductive inference assumption* we want to model. The first implication (A) is not contradictory with human cognitive model that relies on induction, i.e. telling you that *Lucie studies Maths* will not prevent you thinking that she may study *Geometry* or *Algebra*. Implication (B) is related to the fact that inductive inferences made by people are made mostly by considering their understanding of the world. For these two reasons, the implications induced by considering that a non-leaf annotation is imprecise are considered acceptable and mandatory for modelling the *inductive inference assumption* considering the facts: (i) we cannot know which annotation is imprecise and (ii) only a partial representation of the domain is modelled into the ontology.

Our modelling strategy to represent the imprecision of concepts is to associate to each concept a subset of a finite set that can be used to represent the precise answers/annotations. This will be illustrated through an example. Each precise answer corresponds to a precise concept (leaf) defined into the considered finite set. An imprecise concept will be represented using the representations of the precise concepts it subsumes. As an example, the imprecise concept *Maths* will therefore be represented by $\Omega_{Maths} = \{\omega_{M1}, \ldots, \omega_{Mk}\}$ where k is the number of *Maths* descendants that are precise (part of the set of leaves). In our book taxonomy (Fig. 1) $k = 2$, $\Omega_{Maths} = \{\omega_{Algebra}, \omega_{Geometry}\}$. In the following ω will refer to a leaf that corresponds to a precise concept. Therefore,

by considering this setting, the only information that can be deduced from the answer *Lucie studies Maths* to the question *What does Lucie study?* is that: $\omega \in \Omega_{Maths}$.

Let generalize this idea for a taxonomy $O = (\preceq, C)$. We associate each concept in $leaves$ to an element in a finite set denoted $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. For each concept $c \in C$, Ω_c denotes the subset of Ω that only contains the elements associated to leaves in $D(c) \cap leaves$. Then we can generalise the first association between $leaves$ and Ω to an association between C and 2^Ω (the power set of Ω). To each concept $c \in C$ is associated $\Omega_c \in 2^\Omega$. To avoid situations where $\Omega_c = \Omega_{c'}$ with $c \neq c'$, a fictional element ω_c is added to Ω for each concept c that is not a leaf – the representation of c is therefore defined such as $\Omega_c = \bigcup_{x \in D(c)} \{\omega_x\}$. Otherwise stated, we define a specific expression of the function ρ defined in the framework of [6] to characterize the representation of a concept. This representation will be used to manipulate the information carried by a concept: ρ is defined as follow $\rho : C \rightarrow 2^\Omega$ with Ω built as specified above; for convenience we consider $\rho(c) = \Omega_c$. Figure 2 illustrates how the representation of a concept is built.[1]

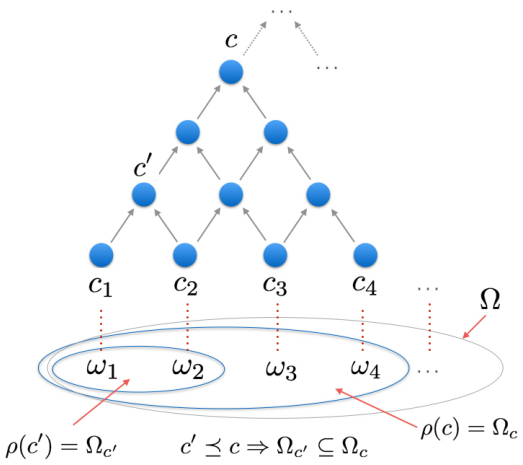

Fig. 2. Illustration of concept representations

Recall that we assume that when a document is annotated by a concept $c \in C$, the annotation is imprecise as soon as c is not a leaf. Any of the leaves subsumed by c is possibly mentioned in the document annotated by c, but the available information is only c. To model this, we reconsider the existing probability framework with evidence theory (also called theory of belief functions). This framework is well-suited to our setting since it introduces a probability distribution defined on 2^Ω. We will now introduce how it can be used to model the *inductive inference assumption*.

[1]Note that, for convenience, the figure does not show the elements of Ω that are added for each non-leaf concept in order to avoid $\Omega_c = \Omega_{c'}$ with $c \neq c'$.

A. Belief functions

The theory of belief functions was introduced by Shafer [7] to model imprecision and uncertainty. It is applied in several domains in which information is provided by imprecise sensors or expert judgements [8]. The most important functions defined in this model are summarized in the following.

Let Ω represents a finite set of elements. A *mass function* m is a probability distribution defined on 2^Ω, it is also called a *basic probability assignment* (*bpa*): $m : 2^\Omega \longrightarrow [0, 1]$. It satisfies the probability condition $\sum_{A \subseteq \Omega} m(A) = 1$; this reflects the convention that one's total belief has measure one. In the initial definition of Shafer, m obeys $m(\emptyset) = 0$, i.e. no belief ought to be committed to the empty set.

Shafer [7] gives the following sense to $m(A), A \subseteq \Omega$: $m(A)$ is the portion of belief that is committed to A and to nothing smaller. Dubois [9] explains this by stating: $m(A)$ is the probability that an agent does not know anything more than $\omega \in A$.

If we try to establish a link with our taxonomy and the instances of our domain, we can say that Ω is used to represent the leaves of the taxonomy. In addition, $m(\Omega_c)$ refers to $\frac{|I^*(c)|}{|I|}$ and more particularly to the probability of observing an instance of c – note that by considering $I^*(c)$ the instances of any descendant of c are excluded.

Elements $A \subseteq \Omega$ such that $m(A) > 0$ are called focal elements and their set is denoted by $\mathbf{F}$. The quantity $m(A)$ measures the belief that is exactly committed to A, not the total belief committed to A. To measure the total belief committed to A, the quantities $m(B)$ for all proper subsets B of A must be added to $m(A)$. It is captured by the *belief function*, $Bel : 2^\Omega \longrightarrow [0, 1]$ which is defined as follow:[2]

$$Bel(A) = \sum_{\substack{B \subseteq \Omega \\ B \subseteq A}} m(B)$$

In our taxonomy $Bel(\Omega_c)$ is the sum of masses of all subsets of Ω_c – these subsets prove Ω_c: if an instance of *Algebra* is observed this proves that an instance of *Maths* is observed. On the other side, in this setting, as well as in frameworks like [2], observing an instance of *Maths* does not inform about what we can say regarding *Algebra*, i.e. according to the *inductive inference assumption*.

In our approach, we propose to deal with this problem through the *plausibility function* as a model of *inductive inference assumption*. A *plausibility function* $Pl : 2^\Omega \longrightarrow [0, 1]$ is defined as follow:

$$Pl(A) = \sum_{\substack{B \subseteq \Omega \\ B \cap A \neq \emptyset}} m(B)$$

[2]As we can see, for a concept c, $Bel(\Omega_c)$ is the quantity defined by Resnik [2] as the probability of observing an instance of concept c denoted $p(c)$: $Bel(\Omega_c) = p(c) = \frac{|I(c)|}{|I|}$. Note that p is not a distribution probability on C since $p(root) = 1$ and elements of C are not considered disjoint for p. Moreover, the distribution bel associated to the measure Bel is neither a probability distribution on Ω since $\sum_{\omega \in \Omega} bel(\omega) = \sum_{\omega \in \Omega} Bel(\{\omega\}) = 1$ does not generally hold, e.g. as soon as a non singleton element of Ω has a mass then $\sum_{\omega \in \Omega} Bel(\{\omega\}) < 1$.

where $Pl(A)$ expresses the extent to which one finds A credible or plausible. As an example, since $\Omega_{Algebra} \subset \Omega_{Maths}$, the mass of Ω_{Maths} is used to compute the plausibility of $\Omega_{Algebra}$. Then the higher the probability of observing concept *Maths*, the more credible observing an instance of *Algebra*.

When focal elements are imprecise, the probability of any event $A \subseteq \Omega$, denoted $Pr(A)$ is imprecise and $Bel(A)$ and $Pl(A)$ represent, respectively, the lower and upper probabilities of event A, that is, $Pr(A) \in [Bel(A), Pl(A)]$.

As our aim is to define an IC using our framework, we have to choose a measure on which the IC will be defined. In the theory of belief function, the pair (Bel, Pl) is the counterpart of the traditional probability Pr. The judicious use of this pair in order to replace Pr is to take them together. However, because the IC is commonly used through an indicator in $\mathbb{R}$, we consider that summarizing the pair (Bel, Pl) by a single value for each concept is the best alternative. In addition of using the single value Bel or Pl, others measures can be proposed. For instance, we could use the probability measure $BetP$ called pignistic probability measure. $BetP$ was proposed by Smets [10] via its associated pignistic probability distribution $BetP_m$ which is derived from the mass function m. $BetP_m$ describes the credal state, it is defined as follow, $\forall \omega \in \Omega$:

$$BetP_m(\omega) = \sum_{\substack{A \subseteq \Omega \\ A \ni \omega}} \frac{m(A)}{|A|}$$

Thus $BetP(A) = \sum_{\omega \in A} BetP_m(\omega)$, $\forall A \subseteq \Omega$. The probability distribution $BetP_m$ can be seen as the weighted sum of uniform probabilities on each focal element where the weights are their masses. In our context, $BetP_m$ is defined for $\omega \in \Omega$, i.e. the elements associated to the leaves. In the definition of $BetP_m$, it is considered that the probability of observing a leaf c such that we observed its ancestor c' is $\frac{1}{|\Omega_{c'}|}$ (uniform probability on $\Omega_{c'}$). All these probabilities are summed on all ancestors of c (only focal elements are considered). Finally, the sum is weighted by the mass of the ancestors of c: $BetP_m(\omega) = \sum_{\Omega_{c'} \ni \omega} \frac{m(\Omega_{c'})}{|\Omega_{c'}|}$.

Like the plausibility measure, the pignistic probability measure takes advantage of the information we possess about an ancestor to deduce information about descendants.

B. Belief functions and bottom-up propagation vs top-down propagation

Let us consider the previous example of a user who owns a 100 book library: 98 are explicitly annotated by *Maths*, 1 is annotated by *Algebra* and 1 is annotated by *Antiquity*. In this case, focal elements are $\mathbf{F} = \{\Omega_{Maths}, \Omega_{Algebra}, \Omega_{Antiquity}\}$. Figure 3 provides *bpa* results on this part of the taxonomy.

As shown in Figure 3, the belief measure computation corresponds to a bottom-up propagation of the instances: if the library contains *Algebra* books it is obvious that it contains *Maths* books (at least *Algebra* books). It corresponds to the classical mechanism used by probabilist approaches as proposed by Resnik. In addition to this bottom-up propagation of instances, plausibility measure also performs a top-down propagation of the instances: we have no evidence that the library of the user contains a book about *Geometry*, but we have some evidence that the library contains *Maths* books. This makes plausible the existence of *Geometry* books in the library and corresponds to the *inductive inference assumption* mechanism.

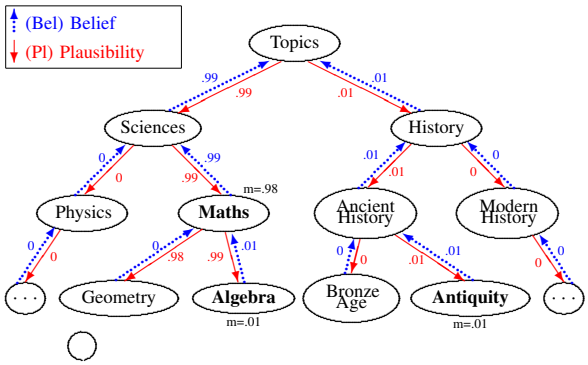

Fig. 3. Mass propagation in taxonomy

C. Using belief functions to compute IC

Let consider a taxonomy $O = (\preceq, C)$, a mass function $m : 2^{\Omega} \rightarrow [0, 1]$, and Bel and Pl, respectively the belief and plausibility functions. As each concept outside *leaves* can take a mass *e.g.* $\exists c \in C \setminus leaves$, such $m(\Omega_c) > 0$, the probability $Pr(\Omega_c)$ is imprecise and we obtain $Pr(\Omega_c) \in [Bel(\Omega_c), Pl(\Omega_c)]$. Considering this setting, several propositions can be stated to compute new ICs for all concept $c \in C$:

1) Bottom-up & Top-down approaches using Plausibility (IC_{Pl}):
$$IC_{Pl}(c) = -log\ Pl(\Omega_c) \quad (4)$$

2) Bottom-up approach using Belief (IC_{Bel}):
$$IC_{Bel}(c) = -log\ Bel(\Omega_c) \quad (5)$$

3) A summarizing convex mean of $Bel(\Omega_c)$ and $Pl(\Omega_c)$, e.g. the arithmetic mean (IC_m with $\alpha = 1/2$):
$$IC_m(c) = -log\left(\alpha Bel(\Omega_c) + (1-\alpha) Pl(\Omega_c)\right)$$

4) Pignistic IC (IC_{Pig}):
$$IC_{Pig}(c) = -log \sum_{\omega \in \Omega_c} BetP_m(\omega) \quad (6)$$

Note that ICs based on Bel, Pl, their mean or $BetP_m$ satisfy the monotony constraint, $x \preceq y \Rightarrow IC(x) > IC(y)$. Table I summarizes the results of IC computations for the example provided in Figure 3 – for each previous proposition $IC_M(c)$ is applied only for measures M with $M(c) \neq 0$. Recall that among the approaches shown in this table only IC_{Bel} does not implement the *inductive inference assumption*. As an example, we can see that IC_{Pl} and IC_{Pig} enable to obtain $IC(Algebra) < IC(Antiquity)$, according to what we wanted to obtain by modelling the aforementioned assumption.

c	IC_{Pl}	IC_{Bel}	$[Bel, Pl]$	IC_{Pig}
Geometry	0.0087			0.31
Algebra	0.004	2	$[0.004, 2]$	0.3
Maths	0.004	0.004	0.004	0.004
Antiquity	2	2	2	2
Topics	0	0	0	0

TABLE I. IC COMPUTATIONS

IV. EVALUATIONS

The aim of our proposal is to define a new IC model that respects the assumption that occurrences of an event must (i) impact the informativeness of both the event and its generalization, but also (ii) expectations of more specific events – this assumption has been detailed through the name *inductive inference assumption*. The fact that IC_{Pl} (Eq. 4) and IC_{Pig} (Eq. 6) correctly model this assumption has been illustrated mathematically in Section III. We have also stressed that existing proposals do not model this assumption and we have underlined that this can be an issue in some specific usage contexts, e.g. to estimate the informativeness of a specific fact – please refer to the example provided in Section II. We therefore consider that integrating this assumption in IC models helps better estimating concept informativeness considering that the usage context agrees with the *inductive inference assumption*. This is not something that has to, or even can, be validated since it is a modelling choice to agree or not with this assumption. Note that, as we did in this paper in order to stress the need of our proposal, because human way of thinking is intuitively tight to inductive procedures, it is easy to build examples for which good IC models will require to take into account this assumption.

Nevertheless, even if the benefits of our proposals compared to existing models have been stressed for several situations, we want to ensure that proposed models do not negatively impact the accuracy of systems that extensively rely on IC models, and that achieve good performances with existing IC models. To this end, we propose to evaluate the impact of the new IC models on the performance of semantic similarity measures. This will be our first evaluation setting. Secondly, we also propose to study the correlations of the IC values estimated by existing and proposed models in a real application setting. This will help us to analyse the effect of considering the *inductive inference assumption* in estimating concept informativeness.

Ideally, we would like to test the performance of our proposals in modelling the *inductive inference assumption*. However, this is something difficult to do since it would require defining a procedure (e.g. metric) that could be used to compare models implementing this assumption. Nevertheless, since (i) we have stressed both the application and the added-value of our proposal – the consideration of the *inductive inference assumption* in IC models –, (ii) no prior model considering this assumption has been proposed in the literature, and (iii) no test exists to evaluate how well this assumption is modelled, this paper will focus on the two aforementioned experiments, that is to say, evaluating the impact of proposed IC modelling on semantic measure accuracy, and studying correlations of existing and proposed IC models. Results can be reproduced using source code and datasets published at https://github.com/sharispe/published_xp.

A. Effect on semantic measure accuracy

A large variety of semantic measures have been proposed to estimate the semantic similarity or relatedness between pairs of words or pairs of concepts. They are largely used in Information Retrieval, Computational linguistics and approximate reasoning to cite a few. Among the best accurate semantic measure models, several proposals extensively depend on accurate IC estimators – existing IC models such as Resnik's, Seco's and Sanchez et al. models have been introduced in Section II. We briefly present two measures that are extensively used in the literature; we will consider them in this experiment. Resnik [2] proposed to estimate the similarity of two concepts c_1 and c_2 defined into a taxonomy using the IC of their Most Informative Common Ancestor $MICA(c_1, c_2)$, i.e. the concept that subsumes both c_1 and c_2 that has the higher IC:

$$\begin{aligned} sim_{Resnik}(c_1, c_2) &= IC(MICA(c_1, c_2)) \\ &= \max_{a \in A(c_1) \cap A(c_2)} IC(a) \end{aligned}$$

One of the drawbacks of Resnik's proposal is that it does not take into account the specificity of compared concepts. To solve this issue, Lin's measure [11] is frequently used:

$$sim_{Lin}(c_1, c_2) = \frac{IC(MICA(c_1, c_2))}{IC(c_1) + IC(c_2)}$$

These two measures can be tuned choosing a specific IC model including the ones that are proposed in this paper, e.g. IC_{Bel}, IC_{Pl}, IC_{Pig}.

Semantic similarity measures are commonly evaluated regarding their capacity to mimic human appreciation of word or concept similarity. To this end, the accuracy of measures is evaluated regarding Pearson's and Spearman's correlations between estimated and expected scores of similarity for a set of pairs of words or concepts. Benchmarks that are used to evaluate semantic measures are therefore composed of pairs of words/concepts for which expected similarity scores are provided – the expected score for a specific pair is generally the averaged value of the several scores of similarity that have been provided by a set of participants for this specific pair. Here we propose to discuss the accuracy of measures considering different IC settings based on existing and proposed models. Three datasets have been used: (i) Rubenstein & Goodenough (RG) [12], 65 pairs of words, (ii) Miller & Charles (MC) [13], 28 pairs of concepts and (iii) SimLex999 (SL) [14], we focused on the 666 pairs of nouns provided by this dataset.

In the experiment, WordNet version 3.1 has been used to obtain the partial order of the concepts we want to compare [15] – only concepts associated to nouns were considered. Aforementioned datasets (RG, ML and SL) provide pairs of words with expected similarity scores. These words are not disambiguated and have to be mapped into WordNet – a same word may refer to several concepts. We used an existing word-to-concept mapping to disambiguate pairs of words that compose RG dataset. For both MC and SL datasets, for each pair, we consider all possible comparisons considering the sets of concepts associated to the two words. In these cases, according to what it is generally done in the literature, only the best similarity score of each pair was considered to compute the correlations. Resnik's IC model, as well as proposed models depend on concept usage analysis. To obtain

statistics about WordNet concept usage we used the *Princeton WordNet Gloss Corpus*[3]. In this experiment, we used both the measures and IC implementations provided by the Semantic Measures Library[4] [16]. IC models introduced in this paper have also been implemented using this library.

Table II presents the results obtained with each dataset/measures. It shows the Pearson and Spearman correlations for each measure configuration for the three datasets (RG, MC and SL). Results are provided for the proposed IC models (Belief, Pignistic and Plausibility) and best state-of-the-art IC models – Seco (Eq. 1), Sanchez (Eq. 2) and Resnik (Eq. 3). We also considered an intrinsic formulation of Resnik IC. This IC is similar to classical Resnik IC by considering a single occurrence of each concept; in the following it is denoted Resnik (i.), i. stands for intrinsic. ICs based on plausibility and pignistic probabilities implement the *inductive inference assumption*, they are associated to the symbol * in the table.

	Pearson		**Spearman**	
	Resnik	Lin	Resnik	Lin
	Rubenstein & Goodenough			
IC Belief	0.478	0.478	0.455	0.432
IC Pignistic*	0.481	0.480	0.455	0.424
IC Plausibility*	0.498	0.498	0.454	0.423
IC Resnik	0.477	0.477	0.468	0.439
IC Resnik (i.)	0.339	0.339	0.320	0.320
IC Seco	0.482	0.480	0.455	0.443
IC Sanchez	0.516	0.514	0.451	0.435
	Miller & Charles			
IC Belief	0.761	0.833	0.756	0.797
IC Pignistic*	0.774	0.843	0.758	0.794
IC Plausibility*	0.827	0.836	0.769	0.791
IC Resnik	0.809	0.838	0.793	0.804
IC Resnik (i.)	0.280	0.281	0.266	0.266
IC Seco	0.808	0.841	0.760	0.808
IC Sanchez	0.836	0.847	0.775	0.795
	SimLex 666			
IC Belief	0.528	0.597	0.531	0.582
IC Pignistic*	0.534	0.594	0.533	0.583
IC Plausibility*	0.527	0.564	0.521	0.565
IC Resnik (i.)	0.114	0.114	0.108	0.108
IC Resnik	0.538	0.601	0.527	0.588
IC Seco	0.482	0.480	0.525	0.592
IC Sanchez	0.541	0.583	0.527	0.583

TABLE II. ACCURACY OF RESNIK'S AND LIN'S SEMANTIC SIMILARITY MEASURES CONSIDERING DIFFERENT IC MODELS

The results that have been obtained on each dataset using Resnik's and Lin's measures are similar. They highlight that proposed IC models (Belief, Pignistic, Plausibility) compete with best state-of-the-art models when evaluated through their effect on semantic measure accuracy. This is not surprising for the IC model based on the Belief function since it is a variant of Resnik's extrinsic IC formulation. However, interestingly, these results show that, in addition to modelling the *inductive inference assumption*, both pignistic and plausibility IC models lead to semantic measure accuracies that are comparable to best efficient state-of-the-art IC models. Since this result could be explained by the fact that these IC behave similarly to accurate state-of-the-art IC models, we propose to evaluate the correlations between the different IC models. Note that, because of the poor accuracies obtained by measures based on Resnik (i.), these results cannot be explained by the fact that IC model selection would have no effect on semantic measure accuracy.

[3]http://wordnet.princeton.edu/glosstag.shtml

[4]http://www.semantic-measures-library.org

B. Correlation between IC models

We have analysed the correlations between the estimations made by the IC models best performing in the previous experiment. To this end, we have computed the IC of each WordNet concept using all models. Table III shows the correlations that have been obtained between pairs of models.

The results underline that existing extrinsic models, i.e. proposed by Sanchez et al. and Seco, have very similar behaviours in the setting of this experiment. We also observe important differences between the IC estimations made by IC models that implement the *inductive inference assumption* (IC Pignistic and Plausibility) and traditional IC models. This result is important since it stresses that, as expected, these two types of IC models indeed behave differently. It therefore means that the good accuracies obtained in the previous experiment were due to the fact that IC models implementing the *inductive inference assumption* can also be used as good IC estimators – at least in the context of semantic similarity assessment. We also observe that the IC based on the pignistic probability has a behaviour that is more similar to the IC belief model than the IC Plausibility model. These results suggest that, according to the theory which ensures that $\forall c \in C, IC_{Pl}(c) \leq IC_{Pig}(c) \leq IC_{Bel}(c)$, the IC based on pignistic probabilities is an interesting solution for designing accurate ICs that model the *inductive inference assumption*, while obtaining IC estimations that are not radically different to those made by traditional models.

We stress that the empirical results that have been obtained for both intrinsic (IC_{Seco}, $IC_{Sanchez}$, $IC_{Resnik(i.)}$) and extrinsic ICs (i.e. IC_{Resnik}, IC_{Pl}, IC_{Pig}, IC_{Bel}) depend on multiple experimental setting variables, e.g. topology of the taxonomy and annotation distribution. Indeed, as an example, if the annotations considered mostly refer to specific/precise concepts, the variations between IC_{Pig} and IC_{Resnik} are expected to be low – because the variations due to the consideration of imprecisions (i.e. top-down propagation of masses) will be low. More experiments have to be performed and analysed in order to critic the degree of generality of the conclusions driven by the results obtained in the experiments presented in this paper. This is a work in progress, additional experiments are currently made to better understand and complete the interesting results obtained in these experiments about new IC models implementing the *inductive inference assumption*

V. CONCLUSION

We have presented new IC models based on the belief function framework; they can be used to estimate the informativeness of concepts defined into a taxonomy by taking into account both topological ordering of concepts and statistics about their usage (e.g. in texts). In particular, through the definition of two extrinsic IC models based on the plausibility function and the pignistic probability, we have presented

IC A	IC B	Pearson	Spearman
Sanchez	Seco	0.903	0.999
Sanchez	Resnik	0.842	0.670
Seco	Resnik	0.900	0.672
Belief	Sanchez	0.547	0.733
Belief	Seco	0.638	0.734
Belief	Resnik	0.860	0.939
Plausibility*	Sanchez	0.195	0.060
Plausibility*	Seco	0.163	0.069
Plausibility*	Resnik	0.227	0.147
Pignistic*	Sanchez	0.462	0.402
Pignistic*	Seco	0.522	0.406
Pignistic*	Resnik	0.697	0.667
Belief	Plausibility*	0.209	0.133
Belief	Pignistic*	0.777	0.661
Plausibility*	Pignistic*	0.424	0.430

TABLE III. CORRELATIONS BETWEEN IC MODELS

innovative IC models implementing the *inductive inference assumption*, i.e. that occurrences of a concept must (i) impact the informativeness of both the concept and its generalization, but also (ii) expectations of the concepts it subsumes. The rationale of considering such a *bring-to-mind* model is that, intuitively, a statement increases the credal state of the statement it entails, e.g. telling you that someone likes books of *Maths* will tend to reinforce you to think that he may like books of *Algebra*. These models have the interesting property to overcome the inability of existing ICs to model this behaviour, despite the fact that it seems to play a central role in human cognition. In addition, by regarding some concepts as imprecise expressions of other concepts, these new IC models propose an original view of the information conveyed by concept occurrences in consideration of concept partial ordering. They also propose solutions to take advantage of pieces of information that were simply excluded by existing models. First empirical analyses based on the study of the impact of IC models on semantic measure accuracy have shown that proposed models compete with best accurate state-of-the-art IC models. For these reasons, we are convinced that expressions of IC modelling the *inductive inference assumption* are relevant for defining models in several settings, e.g. recommendation, information retrieval, computational linguistics. Nevertheless, we stress that the semantics associated to the implementation of the *inductive inference assumption* may not been adapted to all usage contexts – further experiments as well as extensive analyses of IC performances under specific usage contexts (e.g. other that semantic similarity assessment) are therefore required in order to fully understand the benefit of each IC modelling approach; the assumption as well as the implications of proposed IC modelling have been detailed in the paper. Interestingly, our study makes the link between the contributions between areas of research that had, so far, only few connections. Indeed, by underlying the connections between the belief function and Resnik IC, as well as the suitability of plausibility function and pignistic probability for implementing the *inductive inference assumption*, we highlight the interesting impact that the belief function framework may have for areas of research related to IC estimation, semantic measures, and more generally approximate search based on ontologies.

REFERENCES

[1] S. Staab and R. Studer, *Handbook on ontologies*. Springer Science & Business Media, 2010.

[2] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence IJCAI*, vol. 1, 1995, pp. 448–453.

[3] N. Seco, T. Veale, and J. Hayes, "An Intrinsic Information Content Metric for Semantic Similarity in WordNet," in *16th European Conference on Artificial Intelligence*. IOS Press, 2004, pp. 1–5.

[4] D. Sánchez, M. Batet, and D. Isern, "Ontology-based information content computation," *Knowledge-Based Systems*, vol. 24, no. 2, pp. 297–303, Mar. 2011.

[5] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, "Semantic Measures for the Comparison of Units of Language, Concepts or Entities from Text and Knowledge Base Analysis," *ArXiv*, vol. 1310.1285, p. 140, Oct. 2013. [Online]. Available: http://arxiv-web3.library.cornell.edu/abs/1310.1285

[6] S. Harispe, D. Sánchez, S. Ranwez, S. Janaqi, and J. Montmain, "A Framework for Unifying Ontology-based Semantic Similarity Measures: a Study in the Biomedical Domain," *Journal of Biomedical Informatics*, vol. 48, pp. 38–53, 2013.

[7] G. Shafer, *A mathematical theory of evidence (Vol. 1)*. Princeton: Princeton university press, 1976.

[8] A. Imoussaten, J. Montmain, and G. Mauris, "A multicriteria decision support system using a possibility representation for managing inconsistent assessments of experts involved in emergency situations," *International Journal of Intelligent Systems*, vol. 29, no. 1, pp. 50–83, 2014.

[9] D. Dubois and H. Prade, "Formal representations of uncertainty," *Decision-Making Process: Concepts and Methods*, pp. 85–156, 2009.

[10] P. Smets and R. Kennes, "The transferable belief model," *Artificial intelligence*, vol. 66, no. 2, pp. 191–234, 1994.

[11] D. Lin, "An Information-Theoretic Definition of Similarity," in *15th International Conference of Machine Learning*, Madison,WI, 1998, pp. 296–304.

[12] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM*, vol. 8, no. 10, pp. 627–633, Oct. 1965. [Online]. Available: http://portal.acm.org/citation.cfm?id=365628.365657

[13] G. A. Miller and W. G. Charles, "Contextual Correlates of Semantic Similarity," *Language & Cognitive Processes*, vol. 6, no. 1, pp. 1–28, 1991. [Online]. Available: http://eric.ed.gov/ERICWebPortal/recordDetail?accno=EJ431389

[14] F. Hill, R. Reichart, and A. Korhonen, "SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation," Aug. 2014. [Online]. Available: http://arxiv.org/abs/1408.3456

[15] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1998.

[16] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, "The Semantic Measures Library and Toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies," *Bioinformatics*, vol. 30, no. 5, pp. 740–742, 2014.