**ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA**
**CONSIGLIO NAZIONALE DELLE RICERCHE**

**A. Formica**

# CONCEPT SIMILARITY IN FORMAL CONCEPT ANALYSIS: AN INFORMATION CONTENT APPROACH

**R. 643 Settembre 2006**

**Anna Formica** – Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti" del CNR, Viale Manzoni 30 - 00185 Roma, Italy. Email : anna.formica@iasi.cnr.it

**This paper appears in Knowledge-Based Systems 21(1), pp.80-87, Elsevier, 2008**.

ISSN: 1128–3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica, CNR

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.rm.cnr.it
URL: http://www.iasi.rm.cnr.it

## Abstract

*Formal Concept Analysis* (FCA) is revealing interesting in supporting difficult activities that are becoming fundamental in the development of the Semantic Web. Assessing *concept similarity* is one of such activities since it allows the identification of different concepts that are semantically close. In this paper, a method for measuring the similarity of FCA concepts is presented, which is a refinement of a previous proposal of the author. The refinement consists in evaluating the similarity of concept descriptors (attributes) by using the *information content* approach, rather than relying on human domain expertise. The information content approach which has been adopted allows a higher correlation with human judgement than other proposals for evaluating concept similarity in taxonomy defined in the literature.

**Keywords**: *Formal Concept Analysis, Semantic Web, information content, similarity reasoning.*

## 1. Introduction

The purpose of *Formal Concept Analysis* (FCA) [19, 36] is to support the user in analyzing and structuring a domain of interest. Given a domain, a concept in FCA is a pair of sets: a set of objects, which are the instances of the concept in that domain, and a set of attributes, which are the descriptors of the concept.

In the literature different directions are being explored about possible interactions among FCA and Conceptual Modelling [27], Artificial Intelligence [1], Object-Oriented databases [39], and software engineering [35]. Currently, FCA techniques are revealing interesting in supporting difficult activities that are becoming fundamental in the development of the Semantic Web [7, 2, 12]. Assessing concept similarity is one of such activities which is growing in importance within ontology engineering and, in particular, ontology merging and ontology alignment [22]. It requires, in general, human interaction and is, therefore, time-consuming and error-prone.

In this paper, a method for evaluating the similarity of FCA concepts is presented, that is a refinement of a previous proposal of the author [14]. In particular, with respect to the mentioned paper, here the similarity measure is independent of the domain expert knowledge. In fact, the prerequisite of the method presented in [14] is the existence of a predefined domain ontology containing similarity degrees for any pair of concept descriptors (attributes) in the domain. Such similarity degrees are established by a panel of experts in the given domain, according to a consensus system. On the basis of the similarity degrees, a *Similarity Graph* is constructed which allows FCA concept similarity to be evaluated.

In this work, in place of the notion of a *Similarity Graph*, we propose to measure concept descriptor similarity by following the *information content* approach originally introduced by Resnik [29], and successively refined by Lin [24]. Such a method allows us to automatically obtain attribute similarity scores without relying on human domain expertise. In particular, it allows the computation of attribute similarity by using the noun frequencies that are defined in any lexical database for the English language that can be found in the Internet, such as $WordNet$ [37]. Note that in this paper the information content approach of Lin has been chosen since it shows a higher correlation with human judgement with respect to the traditional *edge counting* approach [28], and other proposals for measuring concept descriptor similarity in a taxonomy [20].

A further contribution of this paper consists in comparing the intensional components of FCA concepts according to a method that overcomes the limitations of *Dice*'s function [25], which is often adopted in the literature in order to compare sets of attributes [8, 11].

The paper is organized as follows. In the next section the notion of a *Concept Lattice* is recalled, which is used in FCA to organize and structure concepts. In Section 3, the information content approach is briefly summarized. Successively, in Section 4, the method for evaluating concept similarity in Concept Lattices is presented, followed by the Related Work Section. Section 6 addresses the problem of evaluating the contribution of the work, and Section 7 concludes.

## 2. Formal Concept Analysis

FCA provides a conceptual framework for structuring, analyzing and visualizing data, in order to make them more understandable [36, 19]. In FCA, application domains are organized and structured according to *Concept Lattices*, also referred to as *Galois Graphs*.

### 2.1. Concept Lattices

In FCA a concept is defined within a *context*. A context is a triple $(O,A,R)$, where $O$ and $A$ are two sets of elements called *objects* and *attributes*, respectively, and $R$ is a binary relation between $O$ and $A$. In particular, if $oRa$, for $o \in O$ and $a \in A$, then we say that "the object $o$ has the attribute $a$" or "the attribute $a$ applies to the object $o$".

Given two sets $E$, $I$, such that $E \subseteq O$ and $I \subseteq A$, consider the *dual* sets $E'$ and $I'$, i.e., the sets defined by the attributes applying to all the objects belonging to $E$ and the objects having all the attributes belonging to $I$, respectively, that is:

$$E' = \{a \in A \mid oRa \ \forall \ o \in E\}$$
$$I' = \{o \in O \mid oRa \ \forall \ a \in I\}$$

Then, a *concept* of the context $(O,A,R)$ is a pair $(E,I)$ such that $E \subseteq O$, $I \subseteq A$ and the following conditions hold:

$$E' = I,\ I' = E.$$

The sets $E$ and $I$, representing the concept extensional and intensional components respectively, are referred to as the *extent* and the *intent* of the concept, respectively. Therefore, a concept is a pair of sets where the former consists of precisely those objects which have all attributes from the latter and, conversely, the latter consists of precisely those attributes that apply to all objects from the former.

For instance, consider the example given in [14], concerning a context called *European Cities*. It is recalled below:

$O$ = {Athens, Courmayeur, Innsbruck, London, Paris, Reykjavik, Rome},
$A$ = {Archeological_Site, Beach, Metropolis, Euro, Stream, Skiing_Area}

and $R$ is specified in Table 1, where *Arc*, *Bea*, *Met*, *Eur*, *Str* and *Ski* stand for *Archeological_Site*, *Beach*, *Metropolis*, *Euro*, *Stream*, and *Skiing Area*, respectively.

| | Arc | Bea | Met | Eur | Str | Ski |
|---|---|---|---|---|---|---|
| Athens (A) | x | x | x | x | | |
| Courmayeur (C) | | | | x | | x |
| Innsbruck (I) | | | | x | x | x |
| London (L) | | | x | | x | |
| Paris (P) | | | x | x | x | |
| Reykjavik (Re) | | | x | | | x |
| Rome (Ro) | x | x | x | x | x | |

Table 1: The *European Cities* context

In this context, seven objects are present, each corresponding to a European city, and six attributes. A concept of this context is, for instance, the pair:

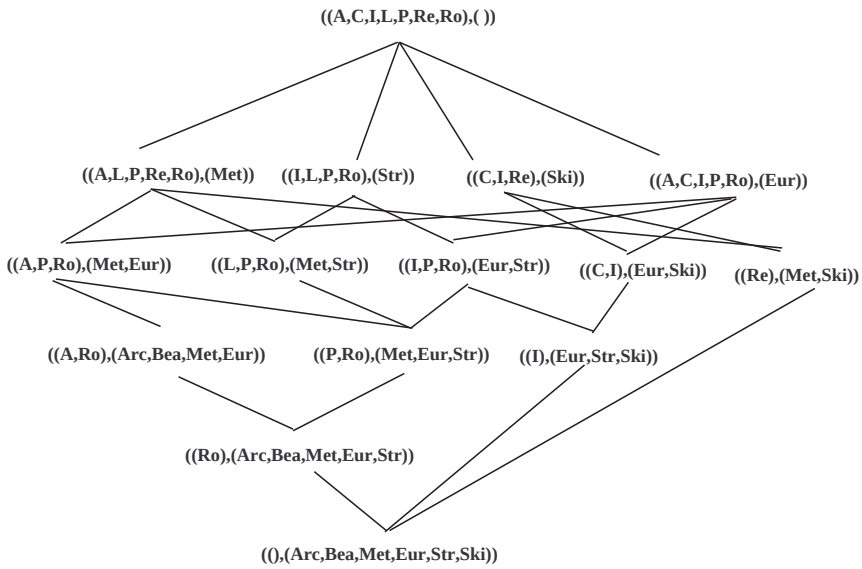


Figure 1: Concept Lattice of the *European Cities* context

((Athens,Paris,Rome), (Metropolis,Euro))

that is, in short form:

((A,P,Ro), (Met,Eur))

In fact, all of *Athens*, *Paris*, and *Rome* have both *Metropolis*, and *Euro* attributes, and viceversa *Metropolis*, and *Euro* together apply to no other object than *Athens*, *Paris*, and *Rome*. Intuitively, it is possible to say that concepts correspond to maximal rectangles of crosses in the context, after appropriate permutations of rows and columns.

Note that, given a context $(O,A,R)$ and two concepts $(E_1,I_1)$ and $(E_2,I_2)$, the following conditions hold:

if $E_1 \subseteq E_2$ then $E_2' \subseteq E_1'$, for $E_1$, $E_2 \subseteq O$
if $I_1 \subseteq I_2$ then $I_2' \subseteq I_1'$, for $I_1$, $I_2 \subseteq A$,

that is, duality implies the opposite set inclusion for both objects and attributes. Therefore, by adding attributes to a concept (i.e., by identifying additional discriminating attributes), the cardinality of its extent decreases, and viceversa, by adding objects to a concept the cardinality of its intent decreases.

Given two concepts $(E_1,I_1)$, $(E_2,I_2)$ of a context $(O,A,R)$, it is possible to establish an *inheritance relation* $(\leq)$ between them according to the following condition:

$(E_1,I_1) \leq (E_2,I_2)$ iff $E_1 \subseteq E_2$ (iff $I_2 \subseteq I_1$).

In particular, $(E_1,I_1)$ is called *subconcept* of $(E_2,I_2)$ and $(E_2,I_2)$ is called *superconcept* of $(E_1,I_1)$.

(Inheritance has been extensively addressed in Conceptual Modelling [10], with particular attention to Object-Oriented databases - for further details see [3, 15].)

Given a context $(O,A,R)$, consider the set of all concepts of this context, indicated as $\mathcal{L}(O,A,R)$. Then:

$$(\mathcal{L}(O,A,R),\leq)$$

is a complete lattice called *Concept Lattice* (also referred to as *Galois Graph*), i.e., for each subset of concepts, the greatest lower bound and the least upper bound exist [36]. (Note that for lattices over sets with finite cardinality, the notions of complete lattice and lattice coincide [9].)

For instance, the Concept Lattice that can be constructed from the context of Table 1 is shown in Figure 1. Note that, nodes are labeled with the concepts of the context, and arcs are established among the nodes whose associated concepts are in $\leq$ relation. The Concept Lattice has also two special nodes, the maximum and minimum nodes (labeled with $\top$ and $\bot$, respectively). The maximum and the minimum group all the objects and the attributes of the context, respectively.

## 3. Information Content Similarity

The notion of *information content similarity* allows similarity of concept intents (attributes) to be computed. This notion is based on the definition of *semantic similarity*, previously introduced in [29], and successively refined in [24]. Before recalling this approach, we need to introduce the notion of a *lexical database for the English nouns* and the related *weighted ISA hierarchy*.

**Definition 3.1. [Lexical database for the English nouns]** *A* lexical database for the English nouns $\mathcal{E}$ *is a 4-uple (N, f(N), R, SynSet), where N is a set of nouns, each associated with a natural language definition, f(N) is a function from N to the positive integers, associating frequencies with nouns, R is a set of relationships on N (such as ISA, PartOf, etc..), and SynSet is the set of sets of nouns of N that are synonyms.*

For instance, consider the $WordNet$ lexical database for the English language [37]. Besides the English nouns, it contains verbs, adjectives, and adverbs, each associated with the related natural language definition and frequency. Nouns are organized essentially according to the $ISA$ and $PartOf$ relationships, and for each noun, a set of synonyms is given ($SynSet$). Therefore, WordNet is also a lexical database for the English nouns, according to the definition above.

Since the information content approach has been conceived for ISA hierarchies, below our attention will focus on the ISA relationship. In particular, the notion of a *weighted ISA hierarchy* derived from a lexical database is introduced. It is based on the notion of *probability* of a concept noun $n$, $p(n)$, which is defined as:

$p(n) = \frac{freq(n)}{M}$

where $freq(n)$ is the *frequency* of $n$ estimated using noun frequencies from large text corpora, as for instance the *Brown Corpus of American English* [17], and $M$ is the total number of observed instances of nouns in the corpus.

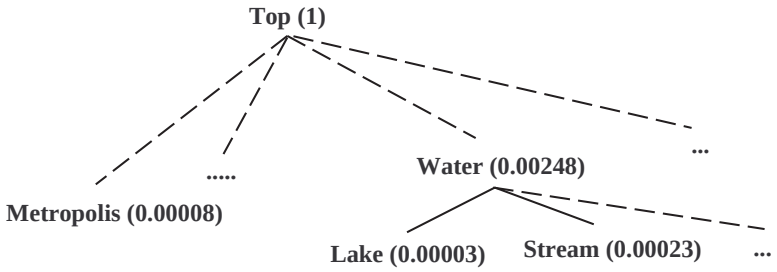


Figure 2: A fragment of the WordNet weighted ISA hierarchy

**Definition 3.2.** **[Weighted ISA hierarchy]** *Given a lexical database for the English nouns $\mathcal{E}$, consider the ISA hierarchy as defined in $\mathcal{E}$. For each node (noun) $n$ of such a hierarchy, consider the probability $p(n)$ that an instance belongs to the concept noun $n$. Furthermore, assume that the ISA hierarchy has a unique* Top *node - the most abstract concept noun - such that $p(Top) = 1$. Such a hierarchy will be indicated as $\mathcal{H}_{\mathcal{E}}$ and will be referred to as the* weighted ISA hierarchy *derived from $\mathcal{E}$.*

In this paper probabilities have been assigned according to the *SemCor* project [13], which labels subsections of the *Brown Corpus* to senses in the *WordNet* lexicon.

Below the definitions of *Water*, *Lake*, *Stream*, and *Metropolis*, and their frequencies (the number in parenthesis), are given:

*(219) Water – the part of the earth's surface covered with water (such as a river or lake or ocean);*
*(3) Lake – a body of (usually fresh) water surrounded by land;*
*(20) Stream – a natural body of running water flowing on or under the earth;*
*(7) Metropolis – a large and densely populated urban area; may include several independent administrative districts;*

A fragment of the weighted ISA hierarchy derived from WordNet is shown in Figure 2 (note that dotted lines stand for undirected ISA links).

With regard to the *SynSet*, by focusing on the sets of synonyms of WordNet that are relevant to our example, the following holds:

$SynSet = \{...,$
$\{Metropolis, City, Urban\_area\},$
$\{Stream, Watercourse\},$
$\{Water, Body\_of\_water\},$
$...\}$

Once probabilities have been associated with nouns, the starting assumption of the approach is that the *information content* of a noun $n$ is defined as - log$p(n)$, that is, as the probability of a concept noun increases, the informativeness decreases, therefore the more abstract a concept noun, the lower its information content [30].

For instance, consider the weighted ISA hierarchy of Figure 2. $Water$ is a concept noun more abstract than *Lake*, therefore, the probability of the former (0.00248) is greater than the probability of the latter (0.00003). As a result, the information content of $Water$ (i.e., - log(0.00248) = 8.66) is less than the information content of *Lake* (i.e., - log(0.00003) = 14.85).

According to this approach, the similarity of hierarchically organized concept nouns is given by the maximum information content shared by the nouns, that is, the more information two nouns share, the more similar they are. Note that given two nouns, say $n_1$,$n_2$, the maximum information content shared by $n_1$,$n_2$ in the taxonomy is provided by the upper bound of $n_1$,$n_2$ whose information content is maximum (i.e., when defined, the least upper bound). Starting from these assumptions, concept nouns similarity according to Lin is defined by the maximum information content shared by the nouns divided by the information contents of the comparing concept nouns. This is formally defined in point 2 of the definition of *information content similarity* below.

**Definition 3.3. [Information content similarity (ics)]** *Given a lexical database for the English nouns $\mathcal{E}$ = (N, f(N), R, SynSet), the derived weighted ISA hierarchy $\mathcal{H}_\mathcal{E}$, and two nouns $n_1$,$n_2 \in N$. The* information content similarity *of $n_1$,$n_2$, indicated as $ics(n_1, n_2)$, is defined as follows:*

1. *if $n_1 = n_2$ or $n_1$,$n_2 \in B_k \in SynSet$, for some $k$:*

$$ics(n_1, n_2) = 1$$

2. *otherwise:*

$$ics(n_1, n_2) = \frac{2 \log p(n')}{\log p(n_1) + \log p(n_2)}$$

*where $n'$ is a concept noun providing the maximum information content shared by $n_1, n_2$, i.e.:*

$$-\log p(n') = \max_{n \in \mathcal{S}(n_1, n_2)} [-\log p(n)]$$

*and $\mathcal{S}(n_1, n_2)$ is the set of concept nouns that are upper bounds of both $n_1$,$n_2$ in the ISA hierarchy.*

In our running example consider *Lake* and *Stream*. Their least upper bound exists in the hierarchy and it is provided by *Water*. Therefore, the following holds:
$ics(Lake,Stream) = \frac{2 \log p(Water)}{\log p(Lake) + \log p(Stream)} = \frac{2 * 8.66}{14.85 + 12.11} = 0.64$.

## 4. Similarity between FCA concepts

In this section the notion of *similarity* ($Sim$) between FCA concepts is introduced. Note that in FCA the concept intent is represented by a set of attributes. Therefore, in the following, in place of concept nouns we will refer to attributes. The comparison of the concept intents presented below has been inspired by the *maximum weighted matching* problem in bipartite graphs, that can be solved in polynomial time [18]. For a formal presentation of the approach, please refer to [16, 14]. Informally, it is illustrated below.

Consider a lexical database for the English nouns $\mathcal{E}$, and two concepts ($E_1$,$I_1$) and ($E_2$,$I_2$) not necessarily belonging to the same context. Let a *candidate set of pairs* be a subset of $I_1 \times I_2$ such that there are no two pairs in the set sharing an element. For instance, assume that $I_1$ and $I_2$ represent a set of boys and a set of girls, respectively, a candidate set of pairs defines a possible set of marriages (when polygamy is not allowed) [18]. Then, within all possible candidate sets of pairs, consider the set such that the sum of the *ics* of the pairs of attributes is maximum. Such a maximum will be indicated as $\mathcal{M}(I_1, I_2)$.

For instance in our running example, assume $I_1 = \{Eur, Str, Ski\}$, and $I_2 = \{Arc, Bea, Met, Eur, Str\}$. Within all possible sets of pairs of attributes that can be formed with $I_1$ and $I_2$ as described above, a set of pairs with maximum sum is the following:

$\{(Eur, Eur),(Str, Str),(Ski, Arc)\}$

since $ics(Eur, Eur) = ics(Str, Str) = 1$, and $ics(Ski, Arc) = 0$. Of course, also the following sets provide the maximum sum:

$\{(Eur, Eur),(Str, Str),(Ski, Bea)\}$

$\{(Eur, Eur),(Str, Str),(Ski, Met)\}$

since:

$ics(Ski, Bea) = ics(Ski, Met) = 0$.

Below the notion of similarity between FCA concepts is presented. It is essentially given by the weighted average between the cardinality of the intersection of the extents of the concepts and the maximum sum $\mathcal{M}(I_1, I_2)$ above.

**Definition 4.1. [Concept similarity (Sim)]** Consider a lexical database for the English nouns $\mathcal{E}$, and two FCA concepts ($E_1$,$I_1$) and ($E_2$,$I_2$) of the same (or different) context(s). Then, the *concept similarity* ($Sim$) between ($E_1$,$I_1$) ($E_2$,$I_2$), $Sim((E_1, I_1), (E_2, I_2))$, is defined as follows:

$$Sim((E_1, I_1), (E_2, I_2)) = \frac{\mid (E_1 \cap E_2) \mid}{r} * w + \frac{\mathcal{M}(I_1, I_2)}{m} * (1 - w)$$

where $\mathcal{M}(I_1, I_2)$ is defined as above, $r$, $m$ are the greatest between the cardinalities of the sets $E_1$,$E_2$, and $I_1$,$I_2$, respectively. Finally $w$ is a weight such that $0 \leq w \leq 1$, that can be established by the user to enrich the flexibility of the method. □

Note that $Sim$ is always a value between zero and one and, for any pair of concepts $(E_1, I_1)$, $(E_2, I_2)$, $Sim((E_1, I_1), (E_2, I_2)) = Sim((E_2, I_2), (E_1, I_1))$.

Consider our running example, and assume $w = \frac{1}{2}$. Let us start by evaluating the similarity of two sibling concepts of the Concept Lattice of Figure 1, namely $((A, Ro), (Arc, Bea, Met, Eur))$, and $((P, Ro), (Met, Eur, Str))$. Since:

$ics(Met, Met) = ics(Eur, Eur) = 1$ and
$ics(Str, Bea) = ics(Str, Arc) = 0$

and $r = 2$, $m = 4$, the following holds:

$Sim[((A, Ro), (Arc, Bea, Met, Eur)), ((P, Ro), (Met, Eur, Str))] =$
$\frac{1}{2} * \frac{1}{2} + \frac{2}{4} * (1 - \frac{1}{2}) = 0.50$

Let us now analyze the similarity between a concept and one of its parents in the Concept Lattice. For instance, consider the concept $((P, Ro), (Met, Eur, Str))$, and the parent $((L, P, Ro), (Met, Str))$. The following holds:

$Sim[((P, Ro), (Met, Eur, Str)), ((L, P, Ro), (Met, Str))] =$
$\frac{2}{3} * \frac{1}{2} + \frac{2}{3} * (1 - \frac{1}{2}) = 0.67$

As also shown in [14], similarity decreases in the case of concepts that are not directly related. For instance, consider again the concept $((P, Ro), (Met, Eur, Str))$ and $((A, C, I, P, Ro), (Eur))$. Then:

$$Sim[((P, Ro), (Met, Eur, Str)), ((A, C, I, P, Ro), (Eur))] = \\ \frac{1}{5} * \frac{1}{2} + \frac{1}{3} * (1 - \frac{1}{2}) = 0.27$$

As already mentioned, the fundamental difference of this work with respect to [14] consists in the evaluation of attribute similarity. For instance, consider the following concept, belonging to a different context, say *Skiing Cities*:

$$((I, Kla), (Lak, Eur, Ski))$$

where $Kla$, and $Lak$ stand for *Klagenfurt* and *Lake*, respectively ($I$, $Eur$, and $Ski$ are defined as in the *European Cities* context). The similarity of this concept with, for instance, the concepts $((I), (Eur, Str, Ski))$ of the Concept Lattice of Figure 1 is the following:

$$Sim[((I, Kla), (Lak, Eur, Ski)), ((I), (Eur, Str, Ski))] = \\ \frac{1}{2} * \frac{1}{2} + \frac{2.64}{3} * (1 - \frac{1}{2}) = 0.69$$

since, as shown in Section 3, $ics(Lak, Str) = 0.64$ and, of course, $ics(Eur, Eur) = 1$, and $ics(Ski, Ski) = 1$.

Note that in [14], we had no way to automatically obtain the similarity degree between *Lake* and *Sream*. In fact, in the previous approach of the author, the analysis performed by a panel of experts in the given application domain was needed, establishing axiomatic similarity degrees for attribute pairs. In this proposal the judgement of the domain experts has been replaced by the notion of *ics* that makes use of the lexical databases for the English language available on the Internet.

## 5. Related Work

Evaluating semantic similarity in FCA is a problem that has been marginally addressed in the literature. To our knowledge, the only relevant proposal concerns fuzzy Concept Lattices, as defined in [4, 5, 6], that are a generalization of the theory of Wille, for the modeling of vague (non-crisp) extents and intents of concepts. In particular, in the mentioned papers an important problem related to FCA has been analyzed, i.e., the large number of concepts that can be extracted from data. This problem is generally addressed by using factorization of Concept Lattices and, in [6], an algorithm for computing a factor lattice of a fuzzy Concept Lattice has been proposed. In particular, factorization is made by similarity, and a similarity measure for concepts of fuzzy Concept Lattices has been proposed. According to this method, that has been extensively presented in [4], similarity is first addressed at level of attributes and objects. For instance, in the case of attributes, two attributes $a_1$, $a_2$ are similar if they cannot be separated by any concept, i.e., if for each concept $c$, $a_1$ belongs to the intent of $c$ if and only if also $a_2$ belongs to the intent of $c$ (analogously in the case of objects). The main difference between the Belohlávek's approach and the one proposed in this paper consists in the similarity of the intensional components of concepts. In fact, in this paper similarity of attributes is established by following the information content approach proposed by [24], and on the basis of it, similarity of

concept intents is computed independently of the related extents (as we have seen, according to a re-visitation of the maximum weighted matching problem in bipartite graphs). In other words, the similarity measure defined in [4] has mainly been conceived for Concept Lattices, therefore by taking into account that the intents and extents of concepts are strictly intertwined. In line with [14], the approach proposed in this paper is more oriented to the Semantic Web and domain ontologies where, in general, the intensional components of concepts are emphasized and can be defined without the extensional components. (Note that, as already mentioned, with respect to [14], we have abandoned the notion of a *Similarity Graph*, and the related notion of axiomatic similarity degree, in order to define a method which is independent of the knowledge of the domain experts.)

Regarding the choice of adopting the information content approach, we recall that, in the literature, the natural, time-honored way to evaluate semantic similarity in a taxonomy (as for instance the WordNet taxonomy) is based on the so-called *edge-counting* approach [23, 28] - that is, the shorter the path between nodes, the more similar the concepts associated with the nodes. Unfortunately, this approach relies on the assumption that links in the taxonomy represent uniform distances that, in general, is a characteristic very difficult to be found in real taxonomies. For this reason, the notion of information content has been originally proposed in [29], and successively refined in [24]. It is independent of the path length of the hierarchy and provides a higher correlation with human judgments [20].

It is worth noting that, in the literature, there are many proposals concerning FCA techniques and ontology merging, knowledge discovering, or data mining [21, 2], although without addressing semantic similarity. We recall that ontology merging consists in taking two or more source ontologies and returning a merged ontology based on the given sources. For instance, in [32, 33] a contribution concerning ontology merging by making use of FCA techniques is presented. In particular, in the mentioned papers the *FCA-merge* method has been proposed, that is based on Ganter and Wille's work on FCA and lattice exploration [19]. Given two or more source ontologies, one context is constructed for each of them, by applying natural language processing techniques. Once the contexts have been defined, they are joined and a pruned Concept Lattice is derived, that is manually explored and transformed into the merged ontology by a knowledge engineer. The engineer has to resolve possible conflicts and duplicates, but there is automatic support from the FCA-merge tool which aims at guiding and focusing the engineer's attention on specific parts of the construction process [31].

Finally, we recall [34], where a method for joining domain ontologies and Galois Graphs for knowledge discovering and data mining has been proposed. In particular, ontologies are used for enhancing keyword-based information retrieval, e.g., filtering the keywords describing a document. Galois Lattices can be used to detect correlations within the knowledge discovery process, and/or to build more concise and accurate domain ontologies.

## 6. Evaluation of the Method

In this paper benchmarks and experimental results have not been performed since, due to the inherently different underlying assumptions of the existing proposals, including [4, 6], they risk to have low relevance. However, some considerations have to be done in order to evaluate this proposal in the absence of experimentation.

As mentioned in the Introduction, the method introduced in this paper is a refinement of a previous proposal of the author presented in [14], which allows FCA concept similarity to be measured independently of the domain expert knowledge. In fact, in the mentioned paper the existence of a predefined domain ontology containing similarity degrees for any pair of concept descriptors (in the given application domain) is assumed. Such similarity degrees are axiomatically established by a panel of experts in the domain, according to a consensus system. A first contribution of this paper consists in the replacement of the axiomatic similarity degrees with the information content similarity (*ics*) scores which can be computed without relying on human domain expertise. In fact, the *ics* can be automatically evaluated according to any lexical database for the English language, as for instance WordNet [37].

A second important consideration has to be done about the choice of Lin's approach. With this regard, it is worth mentioning the problem related to "ideal values". In general, ideal values are established according to human judgement and, in the literature, often the similarity scores assigned by human subjects in the Miller&Charles experiments are addressed (where 28 selected pairs of concepts have been analyzed and one score - on a scale of 0 to 4 - has been given for each pair) [26]. By using the Miller&Charles scores, it has already been shown in [24], and also by other authors, see for instance in [20], that Lin's approach shows a higher correlation with human judgement than other methods for evaluating similarity within a taxonomy, e.g., Resnik [29], Wu&Palmer [38], etc... Therefore, a second contribution of this paper consists in evaluating the similarity of attribute names in FCA according to a proposal which provides similarity scores closer to ideal values than other methods defined in the literature.

A further contribution of this paper regards the comparison of the entire intensional components of FCA concepts, i.e., their sets of attributes. With this regard, in the literature the *Dice*'s function is often adopted [25, 8, 11]. In particular, given two concept nouns, say $c_1$ and $c_2$, each described by a set of attributes, say $I(c_1)$ and $I(c_2)$ respectively, their similarity ($Sim_{Dice}(c_1, c_2)$) is defined as follows:

$Sim_{Dice}(c_1, c_2) = \frac{2|A(c_1,c_2)|}{|I(c_1)|+|I(c_2)|}$

where:

$A(c_1, c_2) = \{(a, b) \mid a \in I(c_1), b \in I(c_2), (a, b) \in C \in \mathit{Aff}\}$

*Aff* is the set of sets of concept nouns showing affinity, and $|A(c_1, c_2)|$, $|I(c_1)|$, and $|I(c_2)|$ are the cardinalities of the sets $A(c_1, c_2)$, $I(c_1)$, and $I(c_2)$, respectively. For instance, consider the FCA concepts:

$((I, Kla), (Lak, Eur, Ski))$

$((I), (Eur, Str, Ski))$

addressed at the end of Section 4. According to *Dice*, depending on the affinity of the attributes *Lake* and *Stream*, the following holds[1]:

$Sim_{Dice}((Lak, Eur, Ski), (Eur, Str, Ski)) = \frac{2*2}{3+3} = 0.67,$

in the case of non-affinity of *Lake* and *Stream*, whereas:

$Sim_{Dice}((Lak, Eur, Ski), (Eur, Str, Ski)) = \frac{2*3}{3+3} = 1.00,$

in the case of affinity of *Lake* and *Stream*. Therefore, in the latter case the set $(Lak, Eur, Ski)$ is considered as a synonym set of $(Eur, Str, Ski)$, which in general does not correspond to human judgement. Then, let us consider the former similarity value, i.e. that related to the non-affinity

[1]Note that, since FCA concepts are not labeled, concept labels which are arguments of $Sim_{Dice}$ have been replaced with concept intents.

of *Lake* and *Stream*. The similarity score, indicated as $SimFCA_{Dice}$ below, that is obtained for the above FCA concepts by addressing $Sim_{Dice}$ (rather than $\mathcal{M}$ as defined in Section 4) is:

$SimFCA_{Dice}[((I, Kla), (Lak, Eur, Ski)), ((I), (Eur, Str, Ski))] = 0.59$.

However, also this result is not satisfactory since the assumption that *Lake* and *Stream* have nothing in common does not fit well with our intuition. This is due to the fact that in *Dice* only the cardinality of the set of pairs of attributes showing affinity is considered, whereas the similarity scores of the pairs are not addressed. With this regard, a third contribution of this paper consists in the possibility of evaluating FCA concepts similarity by explicitly addressing the similarity scores of concept attributes, therefore overcoming the limitations of *Dice*. In fact, we have seen that, in the case of the example above, the affinity between *Lake* and *Stream* is evaluated as $ics(Lak, Str) = 0.64$. On the basis of this similarity value, the following holds:

$\mathcal{M}((Lak, Eur, Ski), (Eur, Str, Ski)) = \frac{2.64}{3} = 0.88$,

which leads to the final similarity value (greater than 0.59):

$Sim[((I, Kla), (Lak, Eur, Ski)), ((I), (Eur, Str, Ski))] = 0.69$,

as shown in Section 4.

## 7. Conclusion

In this paper a similarity measure for FCA concepts has been proposed by refining a previous work of the author [14]. In particular, the similarity of the concept intents (the sets of attributes) has been addressed according to the information content approach [24], rather than the *Similarity Graph*, whose definition requires human domain expertise. This approach allows the similarity of concept attributes to be evaluated by making use of any lexical database for the English language available on the Internet. The choice about the information content approach of Lin is due to the higher correlation this method shows with respect to other traditional approaches for measuring concept descriptor similarity in a taxonomy which are defined in the literature [24, 20].

## References

[1] F.Baader, B.Sertkaya; *Applying Formal Concept Analysis to Description Logics*; International Conference on Formal Concept Analysis (ICFCA), pp.261-286, 2004.

[2] M.Bain; *Inductive Construction of Ontologies from Formal Concept Analysis*; Australian Conference on Artificial Intelligence, pp.88-99, 2003.

[3] C.Beeri, A.Formica, M.Missikoff; *Inheritance Hierarchy Design in Object-Oriented Databases*; Data & Knowledge Engineering (DKE), 30(3), pp.191-216, 1999.

[4] R.Belohlávek; *Similarity relations in concept lattices*; J. Log. Comput. 10(6), pp.823-845, 2000.

[5] R.Belohlávek; *Combination of knowledge in fuzzy concept lattices*; Int. Journal of Knowledge-Based Intelligent Engineering Systems 6(1), pp.9-14, 2002.

[6] R.Belohlávek, J.Dvorák, J.Outrata; *Fast factorization of concept lattices by similarity: solution and an open problem*; In Proc. of Concept Lattices and their Applications (CLA), V.Snásel, R.Belohlávek (Eds), Ostrava, Czech Republic, pp.47-57, 2004.

[7] T.Berners-Lee et al.; *The Semantic Web*; Scientific American, May 2001.

[8] D.Bianchini, V.De Antonellis, M.Melchiori; *Capability Matching and Similarity Reasoning in Service Discovery*; M.Missikoff and A.De Nicola (Eds.), Proc. of the Open Interop Workshop on Enterprise Modelling and Ontologies for Interoperability (EMOI-INTEROP), 13-14 June, Porto, Portugal, CEUR-WS.org, 2005.

[9] G.Birkoff; *Lattice Theory*, Amer. Math. Soc. Providence, R.I., 1967.

[10] A.Borgida, J.Mylopoulos, H.K.T.Wong; *Generalization/Specialization as a Basis for Software Specification*; in "On Conceptual Modelling: Perspectives from Artificial Intelligence, Databases and Programming Languages", pp.87-117, Springer Verlag, 1984.

[11] S.Castano, V.De Antonellis, M.G.Fugini, B.Pernici; *Conceptual Schema Analysis: Techniques and Applications*; ACM Transactions on Database Systems, 23(3), 286-332, 1998.

[12] Y.Ding, D.Fensel, M.Klein, B.Omelayenko; *The semantic web: yet another hip?* Data & Knowledge Engineering 41(2-3), pp.205-227, 2002.

[13] C.Fellbaum; *A Semantic Network of English: the Mother of all WordNets*; Computers and the Humanities 32, 209-220, 1998.

[14] A.Formica; *Ontology-based concept similarity in Formal Concept Analysis*; Information Sciences, 176(18), pp.2624-2641, 2006.

[15] A.Formica, H.D.Groger, M.Missikoff; *Object-Oriented Database Schema Analysis and Inheritance Processing: a Graph-Theoretic Approach*; Data & Knowledge Engineering (DKE), 24(2), pp.157-181, 1997.

[16] A.Formica, M.Missikoff; *Concept Similarity in SymOntos: an Enterprise Ontology Management Tool*; The Computer Journal, 45(6), pp.583-594, 2002.

[17] W.N.Francis, H.Kucera; *Frequency Analysis of English Usage: Lexicon and Grammar*; Houghton Mifflin, 1982.

[18] Z.Galil; *Efficient algorithms for finding maximum matching in graphs*; ACM Computing Surveys, 18, pp.23-38, 1986.

[19] B.Ganter, R.Wille; *Formal Concept Analysis: Mathematical Foundations*; Springer, Berlin, 1999.

[20] J. J.Jiang, D. W. Conrath; *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*; The Computing Research Repository (CoRR), cmp-lg/9709008, 1997.

[21] Y.Kalfoglou, S.Dasmahapatra, Y.Chen-Burger; *FCA in Knowledge Technologies: Experiences and Opportunities*; Int. Conference on Formal Concept Analysis (ICFCA), pp.252-260, 2004.

[22] M.Klein; *Combining and relating ontologies: an analysis of problems and solutions*; in WS on Ontologies and Information Sharing, A.Gomez-Perez et Al. (Eds), IJCAI'01, Seattle, USA, 2001.

[23] J.H.Lee, M.H.Kim, Y.J.Lee; *Information Retrieval Based on Conceptual Distance in IS-A Hierarchies*; Journal of Documentation, 49(2), 188-207, 1993.

[24] D.Lin. *An Information-Theoretic Definition of Similarity.* In Proceedings of the International Conference on Machine Learning, Madison, Wisconsin, USA, Morgan Kaufmann, 296–304, 1998.

[25] Y.S.Maarek, D.M.Berry, G.E.Kaiser; *An Information Retrieval Approach For Automatically Constructing Software Libraries*; IEEE Transactions on Software Engineering, 17(8), 800-813, 1991.

[26] G.A.Miller, W.G.Charles; *Contextual correlates of semantic similarity*; Language and Cognitive Processes, 6(1), 1-28, 1991.

[27] U.Priss; *Formal Concept Analysis in Information Science*; Annual Review of Information Science and Technology (ARIST), Preview Volume 40, 2006.

[28] R.Rada, H.Mili, E.Bicknell, M.Blettner; *Development and application of a metric on semantic nets*; IEEE Transactions on Systems, Man, and Cybernetics, 19(1), pp.17-30, 1989.

[29] P.Resnik. *Using Information Content to Evaluate Semantic Similarity in a Taxonomy.* In Proceedings of the International Joint Conference on Artificial Intelligence, Montreal, Quebec, Canada, August 20-25 1995, Morgan Kaufmann, 448–453, 1995.

[30] S.Ross. *A First Course in Probability.* Macmillan, 1976.

[31] G.Stumme; *Ontology Merging with Formal Concept Analysis*; Semantic Interoperability and Integration, Dagstuhl Seminar Proceedings 04391 IBFI, Germany, 2005.

[32] G.Stumme, A.Maedche; *FCA-MERGE: Bottom-Up Merging of Ontologies*; Proc. of International Joint Conference on Artificial Intelligence (IJCAI), Seattle, USA, pp.225-234, 2001.

[33] G.Stumme, R.Taouil, Y.Bastide, N.Pasquier, L.Lakhal; *Computing iceberg concept lattices with Titanic*; Data & Knowledge Engineering 42(2), pp. 189-222, 2002.

[34] L.Szathmary, A.Napoli; *Knowledge organisation and information retrieval using Galois lattices*, in "Workshop on Knowledge Management and Organizational Memories - 16th European Conference on Artificial Intelligence (ECAI), Valencia, Spain", R.Dieng-Kuntz, N.Matta (Eds), pp.73-78, 2004.

[35] P.Tonella; *Formal Concept Analysis in Software Engineering*; Int. Conference on Software Engineering (ICSE), pp.743-744, 2004.

[36] R.Wille; *Restructuring lattice theory: an approach based on hierarchies of concepts*; Sym. on Ordered Sets, I.Rival (Ed), Reidel, Dordrecht, Boston, 1982.

[37] *WordNet 2.1: A lexical database for the english language* http://www.cogsci.princeton.edu/cgi-bin/webwn, 2005.

[38] Z.Wu, M.Palmer; *Verb semantics and lexical selection*; Proc. of the 32nd Annual Meeting of the Associations for Computational Linguistics; June 27-30, Las Cruces, New Mexico, 133-138, 1994.

[39] A.Yahia, L.Lakhal, R.Cicchetti, J.P.Bordat; *iO2 An Algorithmic Method for Building Inheritance Graphs in Object Database Design*; Proc. of Int. Conference on Conceptual Modeling (ER), Cottbus, Germany, October 1996.