ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
CONSIGLIO NAZIONALE DELLE RICERCHE

A. Formica

CONCEPT SIMILARITY IN FUZZY FORMAL CONCEPT ANALYSIS FOR SEMANTIC WEB

R. 09-01 2009

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.

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

This paper presents a method for evaluating concept similarity within *Fuzzy Formal Concept Analysis*. In the perspective of developing the Semantic Web, such a method can be helpful when the digital resources found on the Internet cannot be treated equally and the integration of fuzzy data becomes fundamental for the search and discovery of information in the Web.

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

1. Introduction

Formal Concept Analysis (FCA), defined by Wille [33], can support critical activities that are important for the development of the Semantic Web [7, 23]. One of these activities is represented by *Similarity Reasoning*, i.e., the identification of different concepts that are semantically close, in order to allow users to find information on the Internet more effectively. Evaluating concept similarity is helpful for Semantic Web service discovery [1], definition of query refinement techniques for search engines [32], semantic information retrieval and integration [27, 29], ontology merging, alignment mapping [28, 11, 19], etc... Searching and discovering information in the Web becomes harder in the presence of vague data, or when the user is not sure about what s/he is looking for, or when some information are more relevant than others. For example, keywords extracted from scientific publications can be used to characterize a research area, however treating all keywords equally is inappropriate because some may be more significant than others [30]. Furthermore, often it is difficult to establish whether a document completely belongs to a research area or not. This type of problems can be tackled with fuzzy information. Fuzzy Formal Concept Analysis (FFCA) is a generalization of FCA for modeling uncertainty information [6].

In this paper a measure for evaluating similarity between FFCA concepts is proposed, referred to as $ConSim_f$. It is an extension of a previous proposal of the author for computing similarity between FCA concepts that was conceived for crisp (non-fuzzy) Concept Lattices [14]. In particular, the notion of fuzzy set similarity, referred to as $SetSim_f$, is introduced in order to evaluate similarity of concept extents. Concept intents are compared according to the *information content* approach, originally introduced by [25] and successively refined in [21], which allows a higher correlation with human judgement than traditional approaches [18].

The paper is organized as follows. In Section 2, the basic definitions of fuzzy theory and FFCA are recalled. Successively, in Section 3, the information content approach is introduced and the notion of *information content similarity* (*ics*) is given. In Section 4, the definition of $ConSim_f$ for evaluating concept similarity in FFCA is presented, which is based on *ics*. Finally in Section 5 the related work is given and Section 6 concludes.

2. Basic Definitions

In this section the basic definitions of fuzzy theory are given and, successively, the Fuzzy Formal Concept Analysis is briefly recalled [30].

2.1. Fuzzy Theory

We start by recalling the notion of *fuzzy set*, *fuzzy set intersection* and *fuzzy set union*.

Definition 2.1. [Fuzzy Set] *Given a domain X, a fuzzy set A in X is characterized by a membership function $\mu_A(x)$ which associates each point in X with a real number in the interval [0,1]:*

$$A = \{(x, \mu_A(x)) \mid x \in X\}$$

The value $\mu_A(x)$ represents the "grade of membership" of x in A.

Definition 2.2. [**Fuzzy Set Intersection**] *The intersection of two fuzzy sets A and B, denoted as* $A \cap B$*, with respective membership functions* $\mu_A(x)$*, and* $\mu_B(x)$*, is a fuzzy set whose membership function is defined as:*

$$\mu_{A \cap B}(x) = min(\mu_A(x), \mu_B(x))$$

Definition 2.3. [**Fuzzy Set Union**] *The union of two fuzzy sets A and B, denoted as* $A \cup B$*, with respective membership functions* $\mu_A(x)$*, and* $\mu_B(x)$*, is a fuzzy set whose membership function is defined as:*

$$\mu_{A \cup B}(x) = max(\mu_A(x), \mu_B(x))$$

In line with [30], below the notions of *fuzzy set cardinality* and *fuzzy set similarity* are given.

Definition 2.4. [**Fuzzy Set Cardinality**] *The cardinality of a fuzzy set A in X, denoted as* $\mid A \mid$*, is defined as:*

$$\mid A \mid = \sum_{x \in X} \mu_A(x)$$

Definition 2.5. [**Fuzzy Set Similarity**] *The similarity between two fuzzy sets A and B, denoted as* $SetSim_f$*, is defined as:*

$$SetSim_f(A, B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid}$$

Finally, given a traditional set of items S (crisp set), we denote as $\phi(S)$ a *fuzzy set generated from* S, i.e., $\phi(S)$ is a fuzzy set where each item in S has a membership value in [0,1]. The notion of *fuzzy relation* concludes this subsection.

Definition 2.6. [**Fuzzy Relation**] *Given two crisp sets A, B, a fuzzy relation on A, B is a fuzzy set generated from* $A \times B$*, i.e.,* $\phi(A \times B)$*.*

2.2. Fuzzy Formal Concept Analysis

FCA provides a conceptual framework for structuring, analyzing and visualizing data in order to make them more understandable [33, 17]. In FCA, application domains are organized and structured according to *Concept Lattices*, also referred to as *Galois Graphs*. *Fuzzy FCA* ($FFCA$) incorporates fuzzy logic into FCA in order to represent vague information. Similar to FCA, in FFCA a concept is defined within a *context*.

Definition 2.7. [**Fuzzy Formal Context**] *A fuzzy formal context is a triple* $K = (O, A, R = \phi(O \times A))$*, where O is a set of objects, A is a set of attributes, and R is a fuzzy relation on* $O \times A$*. Each pair* $(o, a) \in R$ *has a membership value* $\mu(o, a)$ *in [0,1] and we say that "the object o has the attribute a" or "the attribute a applies to the object o" with the grade of membership* $\mu(o, a)$*.*

Definition 2.8. [Fuzzy Formal Concept] *Given a fuzzy formal context* $K = (O, A, R = \phi(O \times A))$*, a confidence threshold* T*, and 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*, with a grade of membership greater than or equal to* T*, and the objects having all the attributes belonging to* I*, with a grade of membership greater than or equal to* T*, respectively, that is:*

$E' = \{a \in A \mid \forall o \in E : \mu(o, a) \geq T)\}$
$I' = \{o \in O \mid \forall a \in I : \mu(o, a) \geq T\}$

A fuzzy formal concept of the fuzzy formal context K *with confidence threshold* T *is a pair* $(\phi(E), I)$*,* $E \subseteq O$*,* $I \subseteq A$*, and* $E' = I$*,* $I' = E$*. Each object* $o \in E$ *has a membership value* μ_o *defined as:*

$$\mu_o = \min_{a \in I} \mu(o, a)$$

where $\mu(o, a)$ *is the membership value between the object* o *and the attribute* a*. If* $I = \emptyset$*,* $\mu_o = 1$ *for every* o*.*

The sets E and I, representing the concept extensional and intensional components respectively, are referred to as the *extent* and the *intent* of the fuzzy concept, respectively. Therefore, a fuzzy concept is a pair where the former element is a fuzzy set consisting of precisely those objects having all attributes from the latter, up to a membership threshold and, conversely, the latter is a set consisting of precisely those attributes that apply to all objects from the former, up to a membership threshold.

For instance, consider the example given in [14], concerning the context called *European Cities*. In order to address fuzzy concepts, in this paper a fragment of that example has been selected where, in particular, some of the attributes for which the membership values are necessarily equal to zero or one (for instance the attribute *euro*) have been removed. The fragment is the following:

$O = \{\text{Rome, London, Paris, Athens, Innsbruck}\}$,
$A = \{\text{Archeological\_Site, Beach, Stream}\}$

and R is specified in Table 1, where *Arch*, stands for *Archeological_Site*.

	Arch	Beach	Stream
Rome (R)	1.0	0.7	1.0
London (L)			1.0
Paris (P)	0.6		1.0
Athens (A)	1.0	0.9	0.3
Innsbruck (I)			1.0

Table 1: Fragment of the *European Cities* context

In this context, five objects are present, each corresponding to a European city, and three attributes. According to FCA (i.e., without fuzziness), supposing that all non-null membership values are assumed equal to 1, a concept of this context is for instance:

((R,A), (Arch,Beach,Stream))
where R, A stand for *Rome* and *Athens*, respectively. In the case of FFCA, assume that the threshold is fixed to 0.5. Then, a concept of this context is, for instance, the pair:
(((R,0.7),(A,0.9)), (Arch,Beach))
that is, the objects Rome, and Athens share two attributes, namely *Arch* and *Beach* and, viceversa, both attributes *Arch* and *Beach* apply to the objects R and A, with membership values greater than or equal to 0.5. In particular, each of the objects A and R is associated with the minimum between the grades of membership relating it with *Arch* and *Beach*. Note that, with respect to the previous example without fuzziness, the attribute *Stream* is not present in the concept since the membership value with *Athens* is 0.3 that is less than the threshold.

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

$$(\phi(E_1), I_1) \leq (\phi(E_2), I_2) \text{ iff } \phi(E_1) \subseteq \phi(E_2) \text{ (iff } I_2 \subseteq I_1)$$

where the definition of fuzzy set inclusion can be derived from that of fuzzy set intersection given in Section 2.1. In particular, $(\phi(E_1), I_1)$ is called *subconcept* of $(\phi(E_2), I_2)$ and $(\phi(E_2), I_2)$ is called *superconcept* of $(\phi(E_1), I_1)$. (Inheritance is a well-known notion that has been extensively addressed in Conceptual Modelling, Artificial Intelligence, Databases and Programming Languages [10], and goes beyond the scope of this paper.)

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

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

is the *Fuzzy Concept Lattice* of the given context, i.e., for each subset of concepts, the greatest lower bound and the least upper bound exist [9, 33].

For instance, the Fuzzy Concept Lattice that can be constructed from the context of Table 1 is shown in Figure 1. Note that nodes are labeled with the fuzzy concepts of the context, and arcs are established among the nodes whose associated concepts are in $\leq$ relation. The Fuzzy Concept Lattice has also two special nodes, the maximum and minimum nodes, grouping all the objects and the attributes of the context, respectively. In particular, the membership values of all the objects of the maximum node is equal to one.

3. Information Content Similarity

The notion of *information content similarity* allows similarity of concept intents (attributes) to be computed. It is based on the definition of *semantic similarity* given in [25, 21]. Before recalling this approach, we have to give the notion of a *lexical database for the English nouns*, and the related definition of a *weighted ISA hierarchy* [14].

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.*

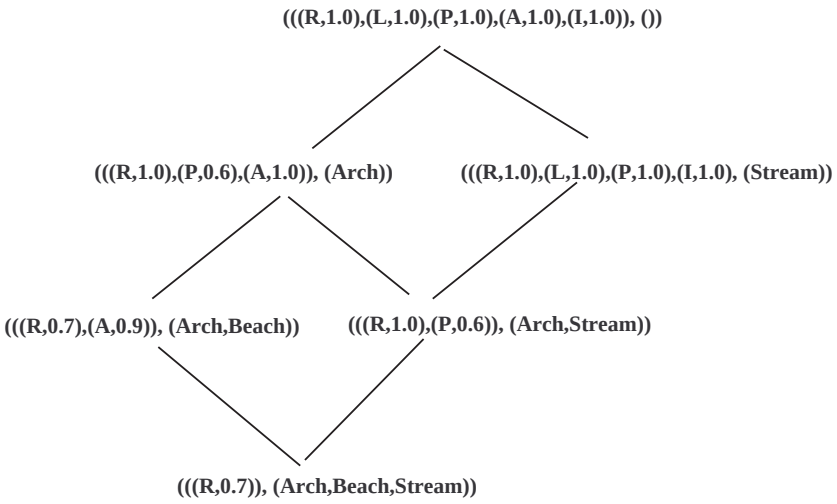

Figure 1: Fuzzy Concept Lattice of (a fragment of) the *European Cities* context

For instance, consider the $WordNet$ lexical database for the English language [34]. 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* [15], and M is the total number of observed instances of nouns in the corpus.

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 $p(n)$ as defined above. 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 [12], which labels subsections of the *Brown Corpus* to senses in the *WordNet* lexicon.

Below the definitions of *Water*, *Lake*, *Stream*, *Beach*, and *Sea*, 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;
(14) Beach – an area of sand sloping down to the water of a sea or lake;

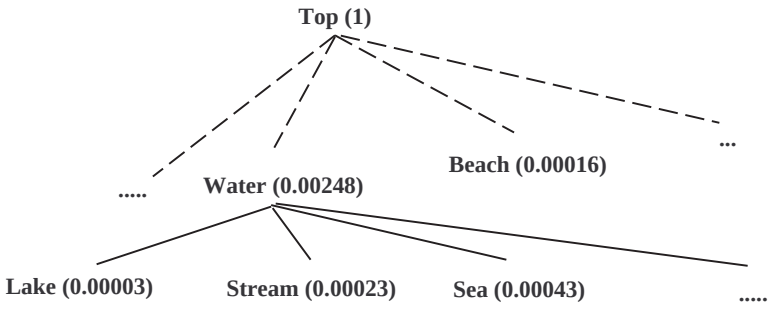

Figure 2: A fragment of the $WordNet$ weighted ISA hierarchy

(38) Sea – a division of an ocean or a large body of salt water partially enclosed by land;

....

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

Just to show a set of sets of synonyms of $WordNet$ that are relevant to our example, we have:
$SynSet = \{...,$
$\{Stream, Watercourse\},$
$\{Water, Body\_of\_water\},$
$...\}$

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 [26].

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 noun similarity according to Lin [21] 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 by 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 FFCA concepts

In this section the notion of similarity between FFCA concepts is introduced. We have seen that in FFCA 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 between concept intents presented below has been inspired by the *maximum weighted matching* problem in bipartite graphs, that can be solved in polynomial time [16]. For a formal presentation of the approach, refer to [13]. Informally, it is illustrated below.

Consider a lexical database for the English nouns $\mathcal{E}$, and two fuzzy concepts ($\phi(E_1)$,I_1) and ($\phi(E_2)$,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). Within all possible candidate sets of pairs, consider (one of) the set(s) such that the sum of the *ics* of the pairs of attributes is maximal. Such a sum will be indicated as $\mathcal{M}(I_1, I_2)$.

For instance in our running example, assume $I_1 = \{Arch, Beach\}$, and $I_2 = \{Arch, Stream\}$. Within all possible sets of pairs of attributes that can be formed with I_1 and I_2 as described above, the set of pairs with maximal sum is the following:
$\{(Arch, Arch)$,$(Beach, Stream)\}$,
since $ics(Arch, Arch) = 1$, and $ics(Beach, Stream) = 0$. Therefore:
$\mathcal{M}((Arch, Beach), (Arch, Stream)) = 1$,
whereas the other possible set of pairs:
$\{(Arch, Stream)$,$(Beach, Arch)\}$
leads to a null value (the *ics* of both the pairs are null).

Below the notion of similarity between FFCA concepts is presented. It is essentially given by the weighted average between the fuzzy set similarity of the concept extents $SetSim_f$, and the maximal sum $\mathcal{M}(I_1, I_2)$ above (up to a normalization factor).

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

$$ConSim_f((\phi(E_1), I_1), (\phi(E_2), I_2)) = SetSim_f(\phi(E_1), \phi(E_2)) * w + \frac{\mathcal{M}(I_1, I_2)}{m} * (1 - w)$$

where $SetSim_f$ is the fuzzy set similarity between the concept extents, $\mathcal{M}(I_1, I_2)$ is defined as above and m is the greatest between the cardinalities of the sets I_1,I_2. Finally w is a weight, such that $0 \leq w \leq 1$, which is established by the domain expert. □

Note that $ConSim_f$ is always a value between zero and one and, for any pair of concepts $(\phi(E_1), I_1)$, $(\phi(E_2), I_2)$:

$$ConSim_f((\phi(E_1), I_1), (\phi(E_2), I_2)) = ConSim_f((\phi(E_2), I_2), (\phi(E_1), I_1)).$$

Consider our running example, and assume $w = \frac{1}{2}$. Let us start by evaluating the similarity of two sibling concepts of the Fuzzy Concept Lattice of Figure 1, for instance:
$(((R, 0.7), (A, 0.9)), (Arch, Beach))$
$(((R, 1.0), (P, 0.6)), (Arch, Stream))$
The fuzzy set similarity $SetSim_f$ of the extents of the concepts above is computed on the basis of the fuzzy set intersection and union as follows (see Section 2):
$((R, 0.7), (A, 0.9)) \cap ((R, 1.0), (P, 0.6)) = 0.7$
$((R, 0.7), (A, 0.9)) \cup ((R, 1.0), (P, 0.6)) = 1.0 + 0.9 + 0.6 = 2.5$
therefore:
$SetSim_f(((R, 0.7), (A, 0.9)), ((R, 1.0), (P, 0.6))) = 0.28$
Furthermore, we have seen that $\mathcal{M}((Arch, Beach), (Arch, Stream)) = 1$. As a result ($m = 2$):
$ConSim_f((((R, 0.7), (A, 0.9)), (Arch, Beach)),$
$(((R, 1.0), (P, 0.6)), (Arch, Stream))) = \frac{1}{2}(0.28 + 0.5) = 0.39.$

Let us now analyze the similarity between two hierarchically related concepts in the Fuzzy Concept Lattice. For instance, consider again the concept:
$(((R, 0.7), (A, 0.9)), (Arch, Beach))$
and its child:
$(((R, 0.7)), (Arch, Beach, Stream))$
With regard to the fuzzy set similarity $SetSim_f$, we have:
$((R, 0.7), (A, 0.9)) \cap ((R, 0.7)) = 0.7$
$((R, 0.7), (A, 0.9)) \cup ((R, 0.7)) = 0.7 + 0.9 = 1.6$
therefore:
$SetSim_f(((R, 0.7), (A, 0.9)), ((R, 0.7))) = 0.44$
Since $\mathcal{M}((Arch, Beach), (Arch, Beach, Stream)) = 2$ and $m = 3$:
$ConSim_f((((R, 0.7), (A, 0.9)), (Arch, Beach)),$
$(((R, 0.7)), (Arch, Beach, Stream))) = \frac{1}{2}(0.44 + 0.67) = 0.56.$

We observe that, in line with [14], the information content approach leads to a fundamental difference with respect to other proposals, including [13]. This point, that will also be discussed

in the Related Work Section, it is illustrated by the following example. Consider the fuzzy concept below, belonging to a different context, say $Architectural\_Cities$:

$(((R, 0.6), (A, 0.8)), (Arch, Sea))$

where R, A stand again for $Rome$ and $Athens$, and an attribute, namely Sea, is present which does not belong to the previous context. The similarity of this concept with, for instance, the concept of the Fuzzy Concept Lattice of Figure 1:

$(((R, 1.0), (P, 0.6)), (Arch, Stream))$

is computed as follows:

$SetSim_f(((R, 0.6), (A, 0.8)), ((R, 1.0), (P, 0.6))) = 0.25.$

Furthermore $\mathcal{M}((Arch, Sea), (Arch, Stream)) = 1 + 0.74$, because:

$ics(Sea, Stream) = \frac{2 \log p(Water)}{\log p(Sea) + \log p(Stream)} = \frac{2*8.66}{11.18+12.11} = 0.74.$

Therefore ($m = 2$):

$ConSim_f((((R, 0.6), (A, 0.8)), (Arch, Sea)),$
$(((R, 1.0), (P, 0.6)), (Arch, Stream))) = \frac{1}{2}(0.25 + 0.87) = 0.56.$

Note that in [13], we had no way to automatically obtain the similarity degree between Sea and $Stream$. In fact, in that paper, 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, as also in [14], the human expertise has been replaced by the notion of *ics* that makes use of lexical databases for the English language available on the Internet (in this example $WordNet$).

5. Related Work

Among the proposals for evaluating concept similarity in FFCA that can be found in the literature, we start by mentioning [30]. In that paper a framework for automatic generation of fuzzy ontologies from uncertainty information has been defined, named FOGA. In FOGA the notion of fuzzy formal concept similarity is essentially based on the similarity of concept extents. Formally, according to the notations and definitions given here, consider two fuzzy formal concepts $(\phi(E_1), I_1)$, $(\phi(E_2), I_2)$. Their similarity, namely Sim, is defined as follows: $Sim((\phi(E_1), I_1), (\phi(E_2), I_2)) = SetSim_f(\phi(E_1), \phi(E_2))$. Therefore, with respect to $ConSim_f$ proposed in this paper, the approach adopted in [30], and analogously in [31], is restrictive since it focuses on the similarity of the concept extents $SetSim_f$, disregarding the similarity of the related intents.

Interesting proposals concerning similarity in Fuzzy Concept Lattices have been defined in [3, 4, 5], mainly to solve the problem related to the large number of concepts that can be extracted from data in a fuzzy concept lattice. In particular, in the mentioned papers similarity is addressed at level of attributes and objects. 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 our proposal similarity of attributes is established according to a re-visitation of the maximum weighted matching problem in bipartite graphs, by following the information content approach proposed by [21], and it is computed independently of the related extents. In other words, the similarity measure defined in [3, 4, 5] has mainly been conceived for Concept Lattices, therefore by taking into account that the intents and extents of concepts are strictly intertwined. Here, in line with [13, 14], the

approach is oriented to the development of the Semantic Web and, in particular, is intended to support activities such as ontology mapping, integration, etc...where, in general, the intensional components of concepts are emphasized and can be defined without the extensional components [2, 8] (however, in this proposal the fuzzy concept extents are not disregarded and play a central role independently of the concept intents). On the other hand, the method proposed in this paper differs from [13, 14] as explained below.

With respect to [14], which focuses on FCA, here fuzzy theory has been addressed and the method has been modified and extended in order to deal with Fuzzy FCA. With respect to [13], in this proposal concept similarity is evaluated independently of human expertise. In fact, in [13] the existence of a predefined domain ontology containing similarity degrees for any pair of attributes, defined in the application domain, is assumed. Such similarity degrees are axiomatically established by a panel of experts in the domain, according to a consensus system. Here, the axiomatic similarity degrees have been replaced with the information content similarity (*ics*) scores which can be computed without relying on human expertise. In fact, the *ics* can be automatically evaluated according to any lexical database for the English language (see for instance, in our running example, the *ics* between $Stream$ and $Lake$).

Note that with respect to other similarity measures proposed in the literature, such *Dice*, or *Cosine* or *Jaccard* [22], a further contribution of this paper consists in the possibility of evaluating FFCA concept similarity by explicitly addressing the similarity scores of concept attributes. In fact, in the mentioned proposals, only the cardinality of the set of pairs of attributes showing affinity is considered, and the similarity scores of the pairs are not addressed. For instance, in our running example, the affinity of the pair ($Lake$,$Stream$) would count 0 or 1 rather than 0.64.

Regarding the choice of Lin's approach, in line with [14], it has been selected since it shows a higher correlation with human judgement (see [21, 18]), than other methods for evaluating similarity within a taxonomy, e.g., Resnik [25], Wu&Palmer [35], etc..., and the traditional time-honored method to evaluate semantic similarity in a taxonomy referred to as *edge-counting* approach [20, 24].

As a final remark, we observe that the evaluation of this proposal has been performed, on one hand, on the basis of theoretic considerations and, on the other hand, on the experimental results existing in the literature. A concrete experimentation has not been given here since, due to the inherently different underlying assumptions of the aforementioned proposals, it risks to have a low relevance. The only comparable proposals are [30, 31] for which, as mentioned above, concept similarity is restricted to concept extent similarity. In fact, the goal of both the mentioned papers is the generation of a fuzzy ontology or a concept hierarchy from uncertainty data rather than the similarity measure itself.

6. Conclusion

Similarity Reasoning, i.e. the identification of different concepts that are semantically close, is becoming fundamental for the retrieval and integration of information within the Semantic Web. FFCA is a generalization of FCA for modeling uncertainty information, which is reveling interesting in supporting critical activities for the development of the Semantic Web, such as ontology mapping, integration, alignment, and Similarity Reasoning [1, 30]. In this paper a measure for evaluating similarity of FFCA concepts is proposed, referred to as $ConSim_f$. It is an extension of a previous proposal of the author for computing similarity between FCA concepts that was conceived for crisp (non-fuzzy) Concept Lattices [14]. With respect to other works

defined in the literature, the contribution of this proposal consists in evaluating FFCA concept similarity as a combination of the similarity of concept extents (fuzzy sets) and concept intents. In particular, concept intents are compared according to the *information content* approach, which allows a higher correlation with human judgement than traditional approaches [21, 18].

References

[1] L.Bai, M.Liu; *A Fuzzy-set based Semantic Similarity Matching Algorithm for Web Service*; Proc. of the IEEE Int. Conference on Services Computing, Volume 2, IEEE Computer Society, 2008.

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

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

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

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

[6] R.Belohlávek, J.Outrata, V.Vychodil; *Fast Factorization by Similarity of Fuzzy Concept Lattices with Hedges* Int. J. of Foundation of Computer Science 19(2), pp.255-269, 2008.

[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] J.Euzenat; *Evaluating ontology alignment methods*; Semantic Interoperability and Integration, Dagstuhl Seminar Proceedings 04391 IBFI, Germany, 2005.

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

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

[14] A.Formica; *Concept similarity in Formal Concept Analysis: an Information Content Approach*; Knowledge-Based Systems, Vol.21, No.1, pp.80-87, Elsevier, 2008.

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

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

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

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

[19] Y.Kalfoglou, W.M.Schorlemmer; *Ontology Mapping: The State of the Art*; Semantic Interoperability and Integration, Dagstuhl Seminar Proceedings 04391 IBFI, Germany, 2005.

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

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

[22] 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), pp.800-813, 1991.

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

[24] 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.

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

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

[27] A.Schwering; *Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey*; Transactions in GIS 12(1), pp. 5-29, 2008.

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

[29] 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.

[30] Q.Thanh Tho, S.Cheung Hui, A.Cheuk, M. Fong, T.Hoang Cao; *Automatic Fuzzy Ontology Generation for Semantic Web*; IEEE Transactions on Knowledge and Data Engineering 18(6), pp.842-856, 2006.

[31] T.Tho Quan, S.Cheung Hui, T.Hoang Cao; *A Fuzzy FCA-based Approach to Conceptual Clustering for Automatic Generation of Concept Hierarchy on Uncertainty Data*; V. Snasel, R. Belohlavek (Eds.), Int. Workshop on Concept Lattices and their Applications (CLA), Ostrava, Czech Republic, September 23-24, pp.1-12, 2004.

[32] DH.Widyantoro, J.Yen; *A fuzzy ontology-based abstract search engine and its user studies* ; Proc. of 10th IEEE International Conference on Fuzzy Systems, December 2-5, Melbourne, Australia, 2001.

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

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

[35] 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, pp.133-138, 1994.