Extension of Fuzzy Galois Connection for Information Retrieval using a Fuzzy Quantifier

C. C. Latiri*, S. Elloumi*, J.P. Chevallet†, A. Jaoua†
*ERPAH Resaerch Team
Computer Science Department
Faculty of Sciences, Tunis
Tunisia
Email: chiraz.latiri@gnet.tn
Email: samir.elloumi@fst.rnu.tn
†MRIM Research Team
Laboratoire CLIPS-IMAG
B.P. 53, 38041 Grenoble Cedex 9
France
Email: jean-pierre.chevallet@imag.fr
‡University of Quatar
Collège of Sciences
Department of Computer Sciences
Doha, PO.Box 2713
jaoua@qu.edu.qa

***Abstract*— In information retrieval, the uncertain implication $D \longrightarrow RQ$ has been different logical status. In this paper, we propose another way to consider this implication in fuzzy context. We introduce a logical formulation based on logical consequence and an evaluation based on an extension of fuzzy Galois connection. The proposed matching function deals with different semantics associated with terms degrees in the query and a fuzzy quantifier.**

I. Introduction

In his 1986 article, Van Rijsbergen [1], [2] has proposed a new IR (Information Retrieval) approach based on logical implication. He suggests to express new matching framework by an implication from a document d to query RQ. He doesn't consider the IR implication $d \longrightarrow QR$ as material implication of the classical logic, but he supposes that the primitive operation for the IR is an uncertain implication.

This approach has inspired some other propositions. In [3], Nie has proposed a logical meta-model where matching is a composition of a direct and reverse implication. He instantiates this model using a fuzzy modal logic. In [4], this model has been extended to a first order fuzzy logic in order to use conceptual graphs as index.

The goal of this paper is to revisit the IR logical model using fuzzy Galois connection as matching framework. In fact, a simple use of the classic Galois connection in IR is illustrated by the two operators of the connection where a first operator f determines the biggest set of terms shared by a set of documents. This operator can be used for indexing in IR. A second operator h permits to retrieve the biggest set of documents indexed by a set of terms which designates the query in this case. So it can be used as a matching function query-document in an IRS (Information Retrieval System).

However, several extensions of the classical Galois connection in fuzzy context has been proposed in literature [5], [6], [7], [8] for fuzzy concept extraction from fuzzy binary relations. But, in the most recent extension [8], the authors have used the implication of Rescher-Gaines that corresponds to the classical case without considering the IR context.

While considering the formal analysis concepts as mathematical background [9], the proposition given in this paper is inspired from the previous work of Pasi [10] who has taken the framework of extended boolean IR model and has adapted it for fuzzy context. The logical model that grounded our work is described in [11]. An attractive idea is to consider that the relevance of a document d can be modelled as a gradual concept related to the truth degree of the implication $t^{RQ} \longrightarrow t^d$ for each term t of the query RQ in a fuzzy context. We examine the possible semantics of this implication, and discuss the more suitable one in an IR application.

This paper is organized as follows: the second section introduces the fuzzy textual data representation. The third section describes basic notions that are useful for understanding the following of the paper. We present in the fourth section the extension of the fuzzy Galois connection. In the fifth section, we describe the new matching framework based on the extension of the fuzzy Galois connection by means a relative fuzzy quantifier and we will conclude then, in the last section, by introducing some perspectives of the approach.

II. A FUZZY REPRESENTATION OF TEXTUAL DATA

First of all, we introduce the definition of term weight in a document as well as the definition of a fuzzy set and a fuzzy binary relation.

Definition 1: The weight P_{ij} of a term t_i in a document d_j, is defined as the representativity factor of the term for this document. The typical measure of the weight is the product of the term frequency by the inverted document frequency, given by the following formula: $P_{ij} = tf_{ij} \times idf$[12].

This weight is often the composition of a local measure to the document and a global measure to the corpus. However, we can reconsider the calculation of this weight using fuzzy set theory. A fuzzy set is defined as follows:

Definition 2: A fuzzy set $\widetilde{F}$ [1] of the discourse universe U is defined by its membership function $\mu_{\widetilde{F}}$: $U \rightarrow [0,1]$, where $\mu_{\widetilde{F}}(u)$, $u \in U$, designates the membership degree of u to the fuzzy set $\widetilde{F}$. The fuzzy set $\widetilde{F}$ is noted as follows: $\widetilde{F} = \{\overset{\mu_{\widetilde{F}}(u_1)}{u_1}, \overset{\mu_{\widetilde{F}}(u_2)}{u_2}, \ldots, \overset{\mu_{\widetilde{F}}(u_n)}{u_n}\}$, where $u_1, ..., u_n \in U$ [13].

So, the choice of the appropriate fuzzy membership function is important to model IR problem. We consider in the following of this paper that the association between a document and the terms which index it can be modeled as a fuzzy binary relation, whatever the used measure in practice to compute the membership degree of the index term to the document. A fuzzy binary relation is defined as follows:

Definition 3: Fuzzy binary relation: Let U and V be two discourse universes. A fuzzy binary relation $\widetilde{R}$ is a fuzzy set defined on the cartesian product $U \times V$. The expression $\mu_{\widetilde{R}}(u, v)$ with $u \in U$ and $v \in V$ expresses the association degree between u and v in the relation $\widetilde{R}$ [13].

In previous IR works, fuzzy representations of textual data have been introduced [14], [15], [16], which their principal objective was to refine and to soften the IR process. The textual database is considered as a fuzzy binary relation denoted by $\widetilde{R}$. Each couple term-document (t_i, d_j) is associated to its weight as shown in table I.

$\widetilde{R}$	t_1	t_2	t_3	t_4
d_1	0.9	0.7	0.6	0.2
d_2	0.7	0.45	0.76	0.8
d_3	0.66	0.88	1.0	0.0
d_4	1.0	0.26	0.6	0.73

TABLE I
FUZZY BINARY REALTION TERM-DOCUMENT

Then, from the fuzzy relation $\widetilde{R}$ in table I, we consider:

1) A finite set D of documents;
2) A finite set T of terms;
3) The couple $\overset{\alpha}{(t_i, d_j)} \in \widetilde{R}$, means that document $d_j \in D$ contains the term $t_i \in T$, with the membership degree α $\in [0, 1]$,which measures the reprsentativity of the term t_i in the document d_j.

In the weighted boolean model [17], the indexation function F takes values in the unit interval $[0, 1]$. In this case, the function F calculates the representativity degree of the concept represented by the term t_i in the document d_j. This value can varied between a nil importance $(F(d_j, t_i) = 0)$ and a full importance $(F(d_j, t_i) = 1)$ [10]. Hence, using fuzzy sets theory, a document $d_j \in D$ can be represented by a fuzzy set of terms, as follows:

$$\widetilde{R}(d_j) = \{(\mu_d(t_i), t_i) \mid t_i \in T\} \qquad (1)$$

The expression (1) indicates that for each term t_i in the set $\widetilde{R}(d_j)$, *a membership degree* is associated $\mu_{d_j}(t_i) \in [0, 1]$. In this case, $\mu_{d_j}(t_i) = F(d_j, t_i)$. It means that the function F is considered as the membership function of the fuzzy set $\widetilde{R}(d_j)$ [14]. A possible definition of the membership function is based on the term weight given in the definition 1. So, the function F is defined by $F(d_j, t_i) = tf_{i,j} \cdot N(idf_j)$, where N is a normalization function to obtain values of the function F in the interval $[0, 1]$ [17].

We will discuss in section V-A, of the different semantics of the degrees associated to terms and what they can bring to IR. We are based in the following of this paper on this fuzzy textual data representation.

III. BASIC NOTIONS

In this section, we present some basic notions related to formal concept analysis [9], fuzzy implications and fuzzy quantifiers, which are fundamental for the rest of this paper.

A. Galois Connection

We define the classical Galois connection while considering the formal classical context $\mathcal{FC} = (D, T, R)$ defined by a finite set of documents D, a finite set of terms T and a binary relation R defined on $D \times T$ [9].

Definition 4: Let R be a binary relation defined on $D \times T$. For two sets A and B, where $A \subseteq D, B \subseteq T$, we define the operators f and h as follows [9]:
$f(A) = \{b \mid \forall a, a \in A \Rightarrow (a, b) \in R\}$
$h(B) = \{a \mid \forall b, b \in B \Rightarrow (a, b) \in R\}$

Proposition 1: The operators f and h define a Galois Connection [9]. The composition $f \circ h$ defines the closure of Galois connection. The operators f and h have the following properties:

- $A_i \subseteq A_j \Rightarrow f(A_i) \supseteq f(A_j)$
- $B_i \subseteq B_j \Rightarrow h(B_i) \supseteq h(B_j)$
- $A_i \subseteq h \circ f(A_i)$ [2] et $B_i \subseteq f \circ h(B_i)$
- $A \subseteq h(B) \Longleftrightarrow B \subseteq f(A)$
- $f = f \circ h \circ f$ et $h = h \circ f \circ h$

B. Fuzzy implications

A fuzzy implication $I = a \rightarrow b$, is a function from $[0, 1] \times [0, 1]$ to $[0, 1]$ that determines the truth degree of the proposition $a \rightarrow b$. In table II, we give the main fuzzy implications.

[1] We use the notation$\,\widetilde{}\,$ for each symbol defined in a fuzzy context.

[2] $h \circ f(A) = h(f(A))$ and $f \circ h(B) = f(h(B))$

Name	Truth Degree	Type
Lukasiewicz $I_{Lucka}(a,b)$	$\min(1, 1-a+b)$	R and S-implication
Gödel $I_{Gö}(a,b)$	$\left\{ \begin{array}{l} 1\ si\ a \leq b \\ b\ otherwise \end{array} \right\}$	R-implication
Goguen $I_G(a,b)$	$\left\{ \begin{array}{l} 1\ si\ a = 0 \\ otherwise \\ \min(1, b/a) \end{array} \right\}$	R-implication
Rescher-Gaines $I_{RG}(a,b)$	1 si $a \leq b$; 0 sinon	R-implication
Kleene-Dienes $I_{KD}(a,b)$	$\max(1-a, b)$	S-implication
Reichenbach $I_{RB}(a,b)$	$(1 - a + a \times b)$	S-implication

TABLE II

FUZZY IMPLICATIONS

The fuzzy implications can be classified in two categories: R-implications and S-implications [18].

C. Fuzzy quantifiers

Zadeh introduces fuzzy quantifiers as linguistic quantifiers to describe intermediate situations between the universal quantifier and the existential quantifier. We distinguish two fuzzy quantifier types: the *relative* quantifiers and the *absolute* ones [19]. In our context of survey, we are interested in the relative quantifiers as *"the most"*.

A relative fuzzy quantifier is defined as a fuzzy set (denoted by Q_{rel}) of the universe discourse U and has the membership function $\mu_{Q_{rel}}$: $U \rightarrow [0,1]$, where $\mu_{Q_{rel}}(u)$ for $u \in U$, is the membership degree of the proportion u to the quantifier Q_{rel} [18].

As example, the relative fuzzy quantifier "*the most*" defined on a set of m terms T, can be represented by the fuzzy set "*most*" as follows:

$$\mu_{most} : T \longrightarrow [0,1]$$

where $\mu_{most}(i) \leq \mu_{most}(i+1)$ for $i = 1..(m-1)$ and $\mu_{most}(m) = 1$.

In the following, we propose to extend the classical Galois connection in IR context, using the fuzzy quantifier "*the most*", taking into account the semantics of membership degrees assigned to the terms of the query. Our objective is to prove the interest of the fuzzy Galois connection and its closure for IR.

IV. FUZZY GALOIS CONNECTION AND IR

The use of the Galois connection in IR supposes a restriction to a boolean model, considering only the conjunction operator. Hence, we propose an extension of the Galois connection in a fuzzy context adapted to IR context. This extension is interesting for IR because it hilights some elements such as the choice of the appropriate fuzzy implication, the use of fuzzy quantifiers, which are original in information retrieval. Our objective is to define a new matching framework based on fuzzy Galois connection.

A. Principle

If we consider the fuzzy representation of textual data illustrated in section II and the extension of fuzzy Galois connection proposed in [8], the fuzzy Galois connection operators $\widetilde{f}$ and $\widetilde{h}$ can be expressed as follows[3]:

$$\widetilde{f}(D) = \{\overset{\alpha_i}{t_i} \mid \alpha_i = \min\{\mu_{\widetilde{R}}(t_i, d), \forall d \in D\}\}$$

$$\widetilde{h}(\widetilde{QR}) = \{d \mid (\mu_{\widetilde{B}}(t_i) \leq \mu_{\widetilde{R}}(t_i, d)), \forall t_i \in \mathcal{T}\} \tag{2}$$

Where:

- $\mathcal{D}$ designates the set of documents of the corpus;
- $\mathcal{T}$ designates the set of indexing terms;
- D is a finit set of documents;
- $\widetilde{QR}$ represents a fuzzy set of terms defined on $\mathcal{T}$.

We notice that the set $\widetilde{QR}$ represents the query in an IR context. So, the operator $\widetilde{h}$ of the fuzzy Galois connection can be defined as the matching function that is going to evaluate the document relevance according to the weighted terms given in a fuzzy query $\widetilde{QR}$. While weighting terms of the query, we specify more restrictions on the retrieved documents. Two types of restrictions are mentioned in literature [20]: *qualitative* restrictions and *quantitative* ones. In our context of survey, we are interested in *qualitative* restrictions that illustrate the case where terms degrees in the query express a criteria that affects the quality of retrieved documents. Thus, degrees translate constraints that must to be satisfied by terms weights in the relation term-document.

B. Discussion

In expression (2), the operator $\widetilde{h}$ is defined using a strict inequality " $\leq$ ", which corresponds exactly to Rescher-Gaines fuzzy implication, illustrated in table II. It is therefore possible to represent the operator $\widetilde{h}$ of expression (2), using Rescher-Gaines fuzzy implication, as follows:

$$\widetilde{h}(\widetilde{QR}) = \{d \mid \forall t_i \in \mathcal{T} \Longrightarrow (\mu_{\widetilde{QR}}(t_i) \underset{I_{RG}}{\longrightarrow} \mu_{\widetilde{R}}(t_i, d)) = 1), \\ \forall d \in \mathcal{D}\} \tag{3}$$

The use of Rescher-Gaines fuzzy implication presents a certain number of limits. Among these, we mention:

- *The noise*: it corresponds to the number of non relevant documents retrieved by the operator $\widetilde{h}$ according to the initial query described by the fuzzy set $\widetilde{QR}$. For example, let us consider the fuzzy relation term-document illustrated in table I, and suppose that $\mu_{\widetilde{QR}}(t_2) = 0.1$.

According to the definition of the operator $\widetilde{h}$ given in expression (3), the truth degree of the implication $\mu_{\widetilde{QR}}(t_2) \underset{I_{RG}}{\longrightarrow} \mu_{\widetilde{R}}(t_2, d_j)$ is equal to 1 for all documents of the relation, i.e. $\mu_{\widetilde{QR}}(t_2) = 0.1 \leq \mu_{\widetilde{R}}(t_2, d_j)_{j=1...4}$. We notice that documents d_3 and d_4 are retrieved although they don't answer to the query

[3] Notations $\widetilde{f}$ and $\widetilde{h}$ proposed for the fuzzy Galois connection, represent an adaptation of the f and h notations of the classical Galois connection, in the fuzzy context.

which asks for the documents not containing the term t_2 with a strong membership degree.

- *The silence:* It corresponds to the number of relevant documents not retrieved by the operator $\widetilde{h}$ according to the query $\widetilde{QR}$. For example, let's consider the binary relation term-document illustrated in table I and let's suppose that $\mu_{\widetilde{QR}}(t_1) = 0.9$. According to the definition of the operator h given in expression (3), the implication $\mu_{\widetilde{QR}}(t_1) \underset{I_{RG}}{\longrightarrow} \mu_{\widetilde{R}}(t_1, d_3)$ deals 0 because $\mu_{\widetilde{QR}}(t_1) = 0.9 > \mu_{\widetilde{R}}(t_1, d_3) = 0.88$, therefore the document d_3 won't be included in the result of $h(\ \widetilde{QR})$ although it is relevant enough since the two values $\mu_{\widetilde{QR}}(t_1) = 0.9$ and $\mu_{\widetilde{R}}(t_1, d_3) = 0.88$ are very near. This proximity can be considered efficiently by using other fuzzy implications.

We notice that Rescher-Gaines implication stays a particular case of fuzzy implication because it rejoins binary case with one binary truth degree equal to 0 or 1.

Furthermore, the use of the operator $\widetilde{h}$ given in expression (3) does not allows to have the relevance degree of the retrieved documents since the set of document D is considered as a crisp one.

These limits justify our orientation toward the use of other fuzzy implications (R-implication or S-implication) to extend the fuzzy Galois connection in IR context by means of a relative fuzzy quantifier.

V. A QUERY MATCHING FRAMEWORK BASED ON EXTENSION OF FUZZY GALOIS CONNECTION

This part shows how to manage a new matching framework in IR using fuzzy Galois connection, extended with a fuzzy implication and a relative fuzzy quantifier. We discuss different possible semantics of fuzzy values in order to express the fuzzy implication $t_i^{RQ} \underset{I_{fuzzy}}{\longrightarrow} t_i^d$ and the interest of using fuzzy quantifier.

A. Semantics of degrees assigned to terms

The proposed matching framework is based on the principle of the fuzzy quotient operator in fuzzy relational databases [21]. In IR context, the division operator is expressed as follows [22]:

$$\mu_{\widetilde{R} \div\ \widetilde{QR}}(d) = \mu_{\widetilde{REL}}(d) = \min_{t \in\ \mathcal{T}}(\mu\ _{\widetilde{QR}}(t) \underset{I_{fuzzy}}{\longrightarrow} \mu_{\widetilde{R}}(d, t)) \qquad (4)$$

where:

- $\widetilde{REL}$ is the fuzzy set of relevant retrieved documents assigned to their relevance degrees, according to the query $\widetilde{QR}$;
- $\underset{I_{fuzzy}}{\longrightarrow}$ is a fuzzy implication;
- and $\mu_{\widetilde{QR}}(t) \underset{I_{fuzzy}}{\longrightarrow} \mu_{\widetilde{R}}(d, t)$ denotes the IR fuzzy implication $t_i^{RQ} \underset{I_{fuzzy}}{\longrightarrow} t_i^d$ which we want to evaluate for each term of the fuzzy query $\widetilde{QR}$.

The implication $t_i^{RQ} \underset{I_{fuzzy}}{\longrightarrow} t_i^d$ shows to what extent of satisfaction, a term t of the query $\widetilde{QR}$ needs to be contained in document d of the fuzzy relation $\widetilde{R}$, where $\mu_{\widetilde{QR}}(t)$ expresses the level of satisfaction required for the term t in the query. It defines the level of term representativity, that the user requires in the retrieved documents indexed by the term t. The document relevance is measured by $\mu_{\widetilde{REL}}(d)$ which is the truth degree of the fuzzy implication.

According to Dubois et al. [22], [21], the proper choice of the fuzzy implication in expression (4) depends on the possible semantic interpretations of the degrees attached to terms. The different meanings are discussed in [21]. In the following, we distinguish two possible meanings for IR:

1) *Semantic of satisfaction*: the level of satisfaction indicates to what extent a term of the query needs to be satisfied. So, this level defines the minimal threshold of the degree attached to each term in the query, which will be considered in the query evaluation process.
2) *Semantic of importance*: a level of importance expresses to what extent a term of the query (with a required level of satisfaction) is important or has a high priority in the global query.

However, in the proposed matching framework, we consider that only semantic of *satisfaction* is associated to the fuzzy binary relation term-document (since $\widetilde{R}$ corresponds to the available information). Both semantics *satisfaction* and *importance* are assigned to the terms weights in the query. Notice that in this paper, these different meanings of weights are only considered to define the operator $\widetilde{h}$ of Galois connection since it is considered as the new fuzzy matching framework.

B. Interest of using fuzzy quantifier in Information Retrieval context

In expression (4), we notice that the document relevance according to the user's query is measured by $\mu_{\widetilde{REL}}(d)$ which is the *minimum* of the truth degree of different fuzzy implications $t_i^{QR} \underset{I_{fuzzy}}{\longrightarrow} t_i^d$ for each term t_i in query $\widetilde{QR}$. The use of Zadeh t-norm *MIN* illustrates the fuzzy conjunction and is justified by the fact that the user aims to retrieve documents indexed by *all* terms in the query $\widetilde{QR}$, which seems too simple in an IR context.

We propose to relax evaluation of the document relevance by means of a typical fuzzy quantifier "*the most*" [19]. The main idea is to be able to retrieve document indexed by "*the most*" terms of the query $\widetilde{QR}$. We notice that Bosc et al. have formulated the related problem by using α-levels cuts on the query $\widetilde{QR}$ [23].

In fact, the use of a fuzzy quantifier "*the most*" corresponds to the situation where there should be only few terms in the query $\widetilde{QR}$ which are not satisfied with a high membership degree in the retrieved documents, from $\widetilde{R}$. Then, a term t in $\widetilde{R}$ with higher level of satisfaction is associated with the higher level of importance in the quantifier "*the most*". Some levels of importance are zero, since "*the most(not all) terms*" are significant in $\widetilde{QR}$. Thus, the concept of "*the most*" can be modelled by $\mu_{most} = 1 - \mu_{I_{most}}$. Using weighted conjunction and fuzzy quantifier "*the most*" [24], the document relevance according to the query $\widetilde{QR}$ is defined by:

$$\mu_{\widetilde{REL}}(d) = \min_{i=1..m}[\max(\mu_{\widetilde{QR}}(t_{\sigma(i)}) \underset{I_{fuzzy}}{\longrightarrow} \tag{5}$$

$$\mu_{\widetilde{R}}(t_{\sigma(i)}, d), 1 - \mu_{I_{most}}(i))], \ \forall \ t \in \mathcal{T}$$

To the fuzzy quantifier "*the most*", we associate a fuzzy set I_{most} defined by: $\mu_{I_{most}}(i) = 1 - \mu_{most}(i-1)$ for $i = 1..m$ and $\mu_{most}(0) = 0$ where:

- The fuzzy set I_{most} expresses the set of ranks of terms that are considered as important in the quantifier "*the most*" and has the following property:

$$\begin{array}{rcl} \mu_{I_{most}}(1) & = & 1 \text{ and } \mu_{I_{most}}(i) \geq \mu_{I_{most}}(i+1) \\ \text{for } i & = & 1..(m-1); \end{array}$$

- When the query $\widetilde{QR}$ contains m terms, the function σ defines a permutation of $(1, ..., m)$ terms of the query $\widetilde{QR}$ with the ordering $\mu_{\widetilde{R}}(t_{\sigma(1)}, d) \geq \mu_{\widetilde{R}}(t_{\sigma(2)}, d) \geq ... \geq \mu_{\widetilde{R}}(t_{\sigma(m)}, d)$ [25].

The contribution of an element $\mu_{\widetilde{R}}(t_{\sigma(i)}, d)$ is fully considered in global evaluation of the relevance if $\mu_{I_{most}}(i) = 1$; namely the term $t_{\sigma(i)}$ is considered as completely important in the quantifier "*the most*" (i.e. it is completely important to satisfy at least i terms). Hence, $\mu_{\widetilde{R}}(t_{\sigma(i)}, d)$ is neglected if $\mu_{I_{most}}(i) = 0$, and $\mu_{\widetilde{R}}(t_{\sigma(i)}, d)$ is considered according to the value of $\mu_{I_{most}}(i)$ if $0 < \mu_{I_{most}}(i) < 1$, that is an intermediate case. In other words, $\mu_{\widetilde{R}}(t_{\sigma(i)}, d)$ is weighted by $\mu_{I_{most}}(i)$ in such a way that the importance degree $\mu_{I_{most}}(i)$ is all the greater as $\mu_{\widetilde{R}}(t_{\sigma(i)}, d)$ is larger. Thus, expression (5) can estimate to what extent the document d is such that $(d, t) \in \widetilde{R}$ holds for term t in the query $\widetilde{QR}$ considered as important under the quantifier "*the most*" rather than for all important terms in $\widetilde{QR}$.

C. Extension of Fuzzy Galois Connection for Information Retrieval

In this approach, we limit ourselves to queries build from conjunction of terms associated with fuzzy degrees under a given semantic. To do not limit to the operator $\min$ which is too simple in an IR context, we promote instead the use the fuzzy quantifier "*the most*". Notice that the operator $\widetilde{h}$ of the extended fuzzy Galois connection defines the new query matching framework. The proposed extension allows to retrieve both relevant document and their relevance degree.

Definition 5: Given $\widetilde{\mathcal{FC}} = (\mathcal{D}, \mathcal{T}, \widetilde{R})$ a textual fuzzy context defined by a finite set of document $\mathcal{D}$, a finite set of terms $\mathcal{T}$ and a fuzzy relation defined on $\mathcal{D} \times \mathcal{T}$. For a fuzzy set of documents $\widetilde{D}$ defined on $\mathcal{D}$ and a fuzzy set of terms $\widetilde{RQ}$ defined on $\mathcal{T}$, $i.e.$ a fuzzy query, the new operators $\widetilde{f}$ and $\widetilde{h}$ of the fuzzy Galois connection are defined as follows:

$$\widetilde{f}(\widetilde{D}) = \{\overset{\alpha}{t} \mid \forall \ d \in \mathcal{D} \Longrightarrow \alpha = \min_i[\max(\mu_{\widetilde{D}}(d_{\sigma(i)})$$

$$\underset{I_{fuzzy}}{\longrightarrow} \mu_{\widetilde{R}}(t, d_{\varphi(i)}), 1 - \mu_{I_{most}}(i))], \forall \ t \in \mathcal{T}\}$$

$$\widetilde{h}(\widetilde{QR}) = \{\overset{\delta}{d} | \ \forall \ t \in T \Longrightarrow \delta = \min_i[\max(\mu_{\widetilde{QR}}(t_{\sigma(i)})$$

$$\underset{I_{fuzzy}}{\longrightarrow} \mu_{\widetilde{R}}(t_{\sigma(i)}, d), 1 - \mu_{I_{most}}(i))], \ \forall d \in \mathcal{D}\}$$

We use the relative fuzzy quantifiers "*the most*" defined in section III-C.

1) Properties of fuzzy Galois connection:

Proposition 2: Let us consider the fuzzy Galois connection $(\widetilde{f}, \widetilde{h})$ illustrated in the definition 5. The following properties are verified for each fuzzy set $\widetilde{D}$ defined on $\mathcal{D}$ and for each fuzzy set $\widetilde{T}$ defined on $\mathcal{T}$, while considering the relative fuzzy quantifier "*the most*" and a fuzzy implication I_{fuzzy}.

Property 1:

- **(P1)** $\widetilde{D}_1 \subseteq \widetilde{D}_2 \Rightarrow \widetilde{f}(\widetilde{D}_1) \supseteq \widetilde{f}(\widetilde{D}_2)$
- **(Q1)** $\widetilde{T}_1 \subseteq \widetilde{T}_2 \Rightarrow \widetilde{h}(\widetilde{T}_1) \supseteq \widetilde{h}(\widetilde{T}_2)$
- **(P2)** $\widetilde{D} \subseteq \widetilde{h} \circ \widetilde{f}(\widetilde{D})$
- **(Q2)** $\widetilde{T} \subseteq \widetilde{f} \circ \widetilde{h}(\widetilde{T})$
- **(P3)** $\widetilde{f}(\widetilde{D}) = \widetilde{f} \circ \widetilde{h} \circ \widetilde{f}(\widetilde{D})$
- **(Q3)** $\widetilde{h}(\widetilde{T}) = \widetilde{h} \circ \widetilde{f} \circ \widetilde{h}(\widetilde{T})$
- **(P4)** $\widetilde{D}_1 \subseteq \widetilde{D}_2 \Rightarrow \widetilde{h} \circ \widetilde{f}(\widetilde{D}_1) \subseteq \widetilde{h} \circ \widetilde{f}(\widetilde{D}_2)$
- **(Q4)** $\widetilde{T}_1 \subseteq \widetilde{T}_2 \Rightarrow \widetilde{f} \circ \widetilde{h}(\widetilde{T}_1) \subseteq \widetilde{f} \circ \widetilde{h}(\widetilde{T}_2)$
- **(P5)** $\widetilde{h} \circ \widetilde{f}(\widetilde{h} \circ \widetilde{f}(\widetilde{D})) = \widetilde{h} \circ \widetilde{f}(\widetilde{D})$
- **(Q5)** $\widetilde{f} \circ \widetilde{h}(\widetilde{f} \circ \widetilde{h}(\widetilde{T})) = \widetilde{f} \circ \widetilde{h}(\widetilde{T})$
- **(PQ6)** $\widetilde{D} \subseteq \widetilde{h}(\widetilde{T}) \Leftrightarrow \widetilde{T} \subseteq \widetilde{f}(\widetilde{D})$

D. Semantic of fuzzy Galois connection Closure in Information Retrieval

In this subsection, we define the semantic of the fuzzy Galois connection closure in Information Retrieval, i.e. $\widetilde{f} \circ \widetilde{h}$. In fact, if we apply the operator $\widetilde{h}$ of the proposed fuzzy Galois connection on a fuzzy set $\widetilde{T}$ of terms, we will retrieve a fuzzy set $\widetilde{D}$ of documents, attached to their respective relevance degrees and which satisfy "*the most*" of the weighted terms in $\widetilde{T}$. In the same way, if we apply the operator $\widetilde{f}$ on a fuzzy set $\widetilde{D}$ of document, we will obtain the fuzzy set $\widetilde{T'}$ of terms associated to their respective weights, which represent the set of weighted terms which index "*the most*" of documents of $\widetilde{D}$. Hence, the closure of the a fuzzy set $\widetilde{T}$ of terms, i.e. $\widetilde{f} \circ \widetilde{h}(\widetilde{T})$ tries to extract a formal fuzzy concept which defines a fuzzy set of terms and a fuzzy set of document strongly linked to each other. In other words, in IR context, the fuzzy Galois connection closure designates a fuzzy set of documents which contains "*the most*" of terms belonging to a fuzzy set of terms and this same fuzzy set of documents represents "*the most*" documents indexed by the fuzzy set of terms. Indeed, we give the definition of *fuzzy reduced concept.*

Definition 6: Given $\widetilde{T}$a fuzzy set of terms defined on $\mathcal{T}$. The fuzzy set $\widetilde{T}$ is called a reduced fuzzy concept, if and only if it is equal to its closure, i.e. $\widetilde{f} \circ \widetilde{h}(\widetilde{T}) = \widetilde{T}$. Thus, $\widetilde{f} \circ \widetilde{h}(\widetilde{T})$ is the minimal fuzzy concept containing $\widetilde{T}$.

E. Semantic query evaluation

We have to distinguish two semantics for the degrees: satisfaction and importance, which leads to using different implications. In the following, we introduce three possible cases [21], [22].

1) First case: The degrees assigned to terms in the query express semantic of satisfaction: When the terms weights reach a given level of satisfaction, this means that for each query term, the user assigns a minimum threshold when he wants to retrieve documents which are indexed by *"the most"* terms in the query. Notice that no priority is imposed among terms in the query. In this way, query $\widetilde{QR}$ is expressed by a fuzzy set $\widetilde{QR}_S$ ($'S'$ expressing the satisfaction semantic). In a more formal way, the relevance of a document is defined by:

$$\mu_{\widetilde{REL}}(d) = \min_i[\max(\mu_{\widetilde{QR}_S}(t_{\sigma(i)}) \underset{R-imp}{\longrightarrow} \mu_{\widetilde{R}}(t_{\sigma(i)}, d), 1 - \mu_{I_{most}}(i))] \tag{6}$$

Where $\mu_{\widetilde{QR}_S}(t_{\sigma(i)})$ expresses the level of representativity required for the term in the document and $\mu_{\widetilde{R}}(t_{\sigma(i)}, d)$ means the membership degree of the term in the fuzzy relation term-document. In order to evaluate the truth degree of this implication, Dubois et al. [25] propose to use an R-implication $\underset{R-imp}{\longrightarrow}$. Expression (6) means to what extent of representativity, a term of the query $\widetilde{QR}_S$ needs to be contained in the relation term-document $\widetilde{R}$. In this case, it is natural that implications give 1, when $\mu_{\widetilde{QR}_S}(t_{\sigma(i)}) \leq \mu_{\widetilde{R}}(t_{\sigma(i)}, d)$. In other words, the level of satisfaction required by the query is reached. This means that the truth degree of the implication $t^{\widetilde{QR}_S}_{\varphi(i)} \underset{R-imp}{\longrightarrow} t^d_{\varphi(i)}$ is evaluated as follows:

$$(t^{\widetilde{QR}_S}_{\varphi(i)} \underset{R-imp}{\longrightarrow} t^d_{\varphi(i)} = 1) \Longleftrightarrow \mu_{\widetilde{QR}}(t_{\sigma(i)}) \leq \mu_{\widetilde{R}}(t_{\sigma(i)}, d)$$

Indeed, Gödel, Goguen et Lukasiewicz [4] implications, which are R-implications, have this characteristic. Otherwise, i.e. $\mu_{\widetilde{QR}_S}(t_{\sigma(i)}) > \mu_{\widetilde{R}}(t_{\sigma(i)}, d)$, the document in the relation $\widetilde{R}$ does not reach the required level of representativity expressed in the query. In this case we distinguish three propositions:

1) Gödel proposition: preserves for the relevance level of the document, the existing the membership degree in the relation $\widetilde{R}$. This means that:

$$(t^{\widetilde{QR}_S}_{\varphi(i)} \underset{R-imp}{\longrightarrow} t^d_{\varphi(i)}) = \mu_{\widetilde{R}}(t_{\sigma(i)}, d)).$$

2) Goguen proposition: keeps a relative degree of satisfaction, that is the actual the membership degree in the relation $\widetilde{R}$ divided by the required degree in the query, namely:

$$(t^{\widetilde{QR}_S}_{\varphi(i)} \underset{R-imp}{\longrightarrow} t^d_{\varphi(i)}) = \frac{\mu_{\widetilde{R}}(t_{\sigma(i)}, d)}{\mu_{\widetilde{QR}_S}(t_{\sigma(i)})}.$$

3) Lukasiewicz proposition: estimates the relevance of the document in terms of the difference$\mu_{\widetilde{QR}_S}(t_{\sigma(i)}) - \mu_{\widetilde{R}}(t_{\sigma(i)}, d)$. This means that the degree of satisfaction can be estimated by:

$$(t^{\widetilde{QR}_S}_{\varphi(i)} \underset{R-imp}{\longrightarrow} t^d_{\varphi(i)}) = 1 - \mu_{\widetilde{QR}_S}(t_{\sigma(i)}) + \mu_{\widetilde{R}}(t_{\sigma(i)}, d).$$

Notice that Gödel implication gives the smallest value of the relevance level of the document. As it is clear in the above discussion, the degree to which a document d satisfies a term t_i in the query $\widetilde{QR}_S$ is determined by comparing $\mu_{\widetilde{QR}_S}(t_i)$ and $\mu_{\widetilde{R}}(t_{\sigma(i)}, d)$.

2) Second case: The degrees assigned to terms in the query express semantic of Importance: In this case, each term of the query has a level of importance not always equal to 1, but all levels of satisfaction which express respective level of representativity are fixed to 1. The query is regarded as a fuzzy set denoted by $\widetilde{QR}_I$ ($'I'$ for importance semantic), which denotes the term priority in the global query. The relevance document is defined as follows:

$$\mu_{\widetilde{REL}}(d) = \min_i[\max(\mu_{\ \widetilde{QR}_I}(t_{\sigma(i)}) \underset{S-imp}{\longrightarrow} \mu_{\widetilde{R}}(t_{\sigma(i)}, d), 1 - \mu_{I_{most}}(i))] \tag{7}$$

Where $\mu_{\ \widetilde{QR}_I}(t_{\sigma(i)})$ expresses the importance level of the term in the query. To evaluate the truth degree of the implication, Dubois et al. [25] suggest the use of S-implication $\underset{S-imp}{\longrightarrow}$. Expression (7) means to what extent an important term in the query $\widetilde{QR}_I$ must be contained in the fuzzy relation term-document $\widetilde{R}$. We assume that $\widetilde{QR}_I$ is normalized (i.e. at least one term of the query has the maximal degree of importance).

Thus, the complete satisfaction of the query can be regarded as demanding that a term, whatever its importance, should be included in the relation $\widetilde{R}$ with the maximal level of representativity; namely:

$$(t^{\widetilde{QR}_I}_{\varphi(i)} \underset{S-imp}{\longrightarrow} t^d_{\varphi(i)} = 1) \Leftrightarrow (\forall\ t_i \in \mathcal{T},\ \mu_{\ \widetilde{QR}_I}(t_{\sigma(i)}) > 0 \Rightarrow \mu_{\widetilde{R}}(t_{\sigma(i)}, d) = 1) \tag{8}$$

when $\mu_{\widetilde{QR}_I}(t_{\sigma(i)}) = 1$ (i.e. the term has the maximal level of importance), document relevance is equal to $\mu_{\widetilde{R}}(t_{\sigma(i)}, d)$. Thus naturally, the relevance level is equal to 0, only if the level of importance is maximal, i.e. $\mu_{\widetilde{QR}_I}(t_i) = 1$, and the membership degree of the term in the document is equal to 0 (i.e. $\mu_{\widetilde{R}}(t_{\sigma(i)}, d) = 0$). Formally we have:

$$(t^{\widetilde{QR}_I}_{\varphi(i)} \underset{S-imp}{\longrightarrow} t^d_{\varphi(i)} = 0) \Longleftrightarrow \mu_{\widetilde{QR}_I}(t_{\sigma(i)}) = 1 \ \ et\ \mu_{\widetilde{R}}(t_{\sigma(i)}, d) = 0 \tag{9}$$

Indeed, any *S-implications* such as Kleene-Dienes and Reinchenbach have the properties (8) and (9). When $\mu_{\ \widetilde{QR}_I}(t_{\sigma(i)}) < 1$, i.e. the term is not very important in the query and has a lesser priority. Thus, even if $\mu_{\widetilde{R}}(t_{\sigma(i)}, d) = 0$, the truth degree of the implication $t^{\widetilde{QR}}_{\varphi(i)} \underset{S-imp}{\longrightarrow} t^d_{\varphi(i)}$ should be strictly positive. Three basic S-implications are considered:

1) If the term can be forgotten to some extent, it leads to use Kleene-Dienes implication illustrated in table II; namely:

$$(t^{\widetilde{QR}_I}_{\varphi(i)} \underset{S-imp}{\longrightarrow} t^d_{\varphi(i)}) = \max(1 - \mu_{\ \widetilde{QR}_I}(t_{\sigma(i)}), \mu_{\widetilde{R}}(t_{\sigma(i)}, d))$$

2) The term can be discounted to some extent. This leads to the using of Reinchenbach implication; namely,

$$(t^{\widetilde{QR}_I}_{\varphi(i)} \underset{S-imp}{\longrightarrow} t^d_{\varphi(i)}) = 1 - \mu_{\ \widetilde{QR}_I}(t_{\sigma(i)}) + \mu_{\ \widetilde{QR}_I}(t_{\sigma(i)}) \times \mu_{\widetilde{R}}(t_{\sigma(i)}, d))$$

[4] Lukasiewicz implication is an R-implication and an S-implication at the same time.

3) The last interpretation is to add the value $1 - \mu_{\widetilde{QR}_I}(t_{\sigma(i)})$ (which represents the amount that is not required) to $\mu_{\widetilde{R}}(t_{\sigma(i)}, d)$; namely:

$$(t_{\varphi(i)}^{\widetilde{QR}_I} \xrightarrow[S-imp]{} t_{\varphi(i)}^{d}) = 1 - \mu_{\widetilde{QR}_I}(t_{\sigma(i)}) + \mu_{\widetilde{R}}(t_{\sigma(i)}, d)).$$

This case corresponds to the use of Lukasiewicz implication.

3) Third case: The degrees assigned to terms in the query express both semantics of Satisfaction and Importance: We consider now the most general case where levels of importance are attached to the terms expressing the satisfying of a specified minimal level of satisfaction in the query. In this case, the level of importance expresses, for example, that is more important to have in the selected document some terms (with a prescribed level of representativity) than to have another term (with its own level of representativity). Using the fuzzy quantifier "*the most*", the relevance document level is defined as follows:

$$\mu_{\widetilde{REL}}(d) = \min[\max(\mu_{\widetilde{QR}_I}(t_{\sigma(i)}) \xrightarrow[S-imp]{} (\mu_{\widetilde{QR}_S}(t_{\sigma(i)}) \xrightarrow[R-imp]{} \mu_{\widetilde{R}}(t_i, d))), 1 - \mu_{I_{mostT}}(i))] \quad (10)$$

Expression (10) estimates to what extent the fuzzy set $\widetilde{QR}_I$ of important terms are included in the fuzzy set ($\widetilde{QR}_S \xrightarrow[R-imp]{} \widetilde{D}$) of terms which sufficiently index the document with respect to the required levels of representativity given by $\widetilde{QR}$. Note that when $\forall\ t_i$ $(\mu_{\widetilde{QR}_I}(t_{\sigma(i)}) = 1$ or when $\forall\ t_i$ $(\mu_{\widetilde{QR}_S}(t_{\sigma(i)}) = 1$, expression (10) is reduced to the two particular cases previously encountered [25].

VI. Conclusion

We have proposed in this paper another way to evaluate the information retrieval implication $D \longrightarrow RQ$ in a fuzzy context. We transform the initial implication into a fuzzy implication and the matching process is modeled by the extension of the fuzzy Galois connection. We want to implement a retrieval engine only based on the computation of this fuzzy Galois connection. Another research area we want to explore is a combination of the direct and reverse implication, as suggested by the initial general logic framework proposed by Nie [3]. As work in progress, we propose to use the proposed extension of the fuzzy Galois connection in textmining to extract fuzzy association rules between terms and perform query expansion using these fuzzy associations between terms.

References

[1] C. V. Rijsbergen, "A non-classical logic for information retrieval," *The Computer Journal*, vol. 29, no. 6, pp. 481–485, 1986.

[2] ——, "A new theorical framework for information retrieval," in *Proceedings of the 1986-ACM Conference on Research and Development in Information Retrieval*, 1986, pp. 194–200.

[3] J. Nie, "Un modèle logique général pour les Systèmes de Recherche d'Informations. Application au prototype RIMRE," Laboratoire de Génie Informatique - IMAG, Université Joseph Fourier, Grenoble I, Thèse de Doctorat, Juillet 1990.

[4] J. Chevallet, "Un modèle logique de Recherche d'Informations appliqué au formalisme des Graphes Conceptuels. Le prototype ELEN et son expérimentation sur un corpus de composants logiciels," Laboratoire de Génie Informatique - IMAG, Université de Joseph Fourier, Grenoble I, Thèse de Doctorat, Novembre 1992.

[5] V. Novak, *Fuzzy sets and their applications*. Adam Higler, 1989.

[6] S. Polland, *Fuzzy-Concepts. Formal Concept Analysis for imprecise data*. Springer Verlag, Berlin, 1996.

[7] R. Belohlavek, "Fuzzy Galois connections," *Math. Logic*, vol. 45, no. 4, pp. 497–504, 1999.

[8] A. Jaoua, S. Elloumi, S. BenYahia, and F. Alvi, "Galois connection in fuzzy binary relations: applications for discovering association rules and decision making," J. Desharnais, M. Frappier, and W. Mccaul, Eds. Methodos Publisher, 2002.

[9] B. Ganter and R. Wille, *Formal Concept Analysis*. Springer-Verlag, Heidelberg, 1999.

[10] G. Pasi, "A logical formulation of the booolean model and of weighted boolean model," in *Proceedings of the Workshop on Logical and Uncertainty Models for Information Systems, London, UK*, 1999, pp. 1–11.

[11] Y. Chiaramella and J. Chevallet, "About Retrieval Models and Logic," *The Computer Journal*, vol. 35, no. 3, pp. 233–242, 1992.

[12] G. Salton and C. Buckely, "Term weighting approaches in automatic text retrieval," in *Information Processing and Management*, May 1988, pp. 513–523.

[13] L. Zadeh, "Fuzzy sets," *Information and Control*, no. 69, pp. 338–353, June 1965.

[14] T. Radecki, "Fuzzy Set Theoretical Approach to Document Retrieval," *Information Processing and Management*, vol. 15, pp. 247–259, 1979.

[15] G. Salton, C. Buckley, and C. Yu, *An Evaluation of Term Dependence Models in Information Retrieval*, 1982.

[16] D. Kraft and D. Buel, *Fuzzy sets and generalized boolean retrieval systems*. Int. J. Man-Machine Studies, 1983.

[17] G. Salton, E. Fox, and H. WU, "Extended Boolean Information Retrieval," *Communications of the ACM*, vol. 26, no. 11, pp. 1022–1036, November 1983.

[18] B. Bouchon-Meunier, *La logique floue et ses applications*. Addison Wesley, 1995.

[19] L. A. Zadeh, "A computational approach to fuzzy quantifiers in natural languages," *Computing and Mathematics with Applications*, no. 9, pp. 149–184, 1983.

[20] E. Viedma, "An information retrieval system with ordinal linguistic queries based on the weighting semantics," in *Proceedings of the 7th International Conference on Information Processing and Management of Uncertainty in knowledge based System, Madrid, Spain*, 2000.

[21] D. Dubois and H. Prade, "Semantics of quotient operators in fuzzy relational databases," *Fuzzy sets and systems*, vol. 78, pp. 89–93, 1996.

[22] ——, "The three semantics of fuzzy sets," *Fuzzy Sets and Systems*, no. 90, pp. 141–150, 1997.

[23] P. Bosc, L. Liétard, and H. Prade, "On fuzzy queries involving fuzzy quantifiers," in *Proceedings of the ECAI Workshop on Uncertainty in Information Systems. Budapest, Hungry*, August 1996, pp. 6–10.

[24] D. Dubois and H. Prade, "Weighted minimum and maximum operations in fuzzy set theory," *Information Sciences*, no. 39, pp. 205–210, 1986.

[25] D. Dubois, M. Nakata, and H. Prade, "Extended division for flexible queries in relational databases," in *Proceedings of the 7th World Congress of the International Fuzzy Systems Associations IFSA '97, Prague*, 1997, pp. 25–29.