2009 IEEE International Conference on Semantic Computing

# An Information Retrieval Model Based On Automatically Learnt Concept Hierarchies

Pawan Goyal[1], Laxmidhar Behera[1,2], T.M.McGinnity[1]
[1]*Intelligent System Research Centre, School of Computing and Intelligent Systems,*
*University of Ulster, UK.*
*goyal-p@email.ulster.ac.uk*, {*l.behera, TM.McGinnity*}*@ulster.ac.uk*
[2]*Department of Electrical Engineering, Indian Institute of Technology, Kanpur, India.*
*lbehera@iitk.ac.in*

***Abstract*—The paper investigates the application of fuzzy logic based concept summarization and formal concept analysis in automatically building concept hierarchies from a text corpora. The context of a term has been modeled using its syntactic relations with the most frequent verbs, which act as attributes. This context information has been used to produce a concept lattice, which retains the concept hierarchies as well as the membership weights of the objects. The concepts within each hierarchy have been summarized using a fuzzy logic based soft least upper bound approach. An information retrieval model is proposed, which uses fuzzy formal concepts to get the relevance degree between the document and the query. Results for ontology evaluation are shown on two domain ontologies.**

***Keywords*-Fuzzy Formal Concept Analysis; Information Retrieval; Ontology Learning**

## I. Introduction

The issues in information retrieval have become more important due to the explosion in availability of documents on the World Wide Web (WWW). The most traditional mechanism for indexing a document is boolean indexing, based on presence or absence of a word in a document. Other traditional mechanisms index documents based on the frequency count of words. The knowledge based fuzzy information search models [1] [2] have shown to improve the efficiency of search systems. A nice survey of the existing literature on 'soft web mining' has been given in [3]. These approaches, however, do not succeed in defining the document content fully because they do not process the concepts in the documents, only the words are processed. An indexing process based on relational ontologies, as introduced by Pereira [4] has been adopted for semantic processing of documents.

Ontologies allow users to define their own vocabulary and the constraints in ontology capture the intended meaning of user-defined vocabulary. Many applications involving text clustering [5], classification [6] and information retrieval [7] have been benefitted by the fact that the ontologies generalize over words. However, developing a domain ontology is an expensive and time-consuming task. Different methods have been proposed in the literature to address the problem of learning ontologies from text. Cimiano [8] has proposed an approach based on Formal Concept Analysis (FCA) following the distributional hypothesis [9]. Navilgi [10] presented Ontolearn to extract domain ontology from available documents, where complex domain terms are semantically interpreted and arranged in a hierarchical fashion. Tho [11] has used the notion of fuzzy formal concept analysis and proposed the Fuzzy Ontology Generation Framework (FOGA). Zouaq [12] presents a semi automatic method to transform textual resources, into domain concept maps, which are transformed into a formal domain ontology.

In our work, fuzzy formal concept analysis has been used. Noun terms have been used as the objects and attributes have been derived by parsing a corpus and extracting verb/object and verb/subject dependencies. For each noun, these verbs are used as attributes for building the formal context and formal concept lattice, such that membership weights of objects within each concept are preserved. Our work differs from that of Cimiano [8] and Tho [11] in that we have used a fuzzy logic based soft least upper bound approach [13] for giving a meaningful representation to the concept nodes in the learnt hierarchy using a general purpose ontology. The formulation for soft least upper bound approach has been modified to take into account the fuzzy membership of the objects within a concept node. The complete algorithm has been proposed for labeling the non-leaf nodes of the hierarchy.

An algorithm for document retrieval has been proposed using the concept hierarchies obtained. The similarity measure between the formal concepts has been used to extract fuzzy propositions, which rank the documents with respect to a user query. The proposed algorithm is a modification to the fuzzy relational ontological model proposed by Pereira [4]. The complete mathematical formulation of the proposed document ranking algorithm has been shown.

The paper has been organized as follows. Section II gives the mathematical preliminaries of fuzzy formal concept analysis. Section III describes the concept summarization approach using soft least upper bound. Section IV describes the information retrieval approach based on fuzzy formal concept analysis. The results obtained are described in

978-0-7695-3800-6/09 $26.00 © 2009 IEEE
DOI 10.1109/ICSC.2009.108

IEEE computer society

section V. The conclusions and discussions are given in section VI.

## II. Formal concept Analysis

Formal concept analysis [14] refers to both an unsupervised machine learning technique and, more broadly, a method of data analysis. The approach takes as input a matrix specifying a set of objects and the properties thereof, called attributes, and finds the "natural" clusters of attributes and objects in the input data, where

- A "natural" object cluster is the set of all objects that share a common subset of attributes, and
- A "natural" property cluster is the set of all attributes shared by one of the natural object clusters.

The family of these concepts obeys the mathematical axioms defining a lattice, and is called a concept lattice. FCA can be seen as a conceptual clustering technique as it also provides intensional descriptions for the abstract concepts or data units it produces. The notations for the fuzzy formal concept analysis [11] are briefly discussed in the subsection II-A.

### A. Fuzzy Formal Concept Analysis

**Definition 1 (Fuzzy formal Context)** A fuzzy formal context is a triple $K = (G, M, I = \psi(G \times M))$, where $G$ is a set of objects, $M$ is a set of attributes, and $I$ is a fuzzy set on domain $G \times M$. Each relation $(g, m) \in I$ has a membership value $\mu(g, m)$ in $[0, 1]$.

Each object $O$ in a fuzzy formal context $K$ is represented by a fuzzy set $\phi(O)$ as

$$\phi(O) = \{A_1(\mu_1), A_2(\mu_2), \ldots, A_m(\mu_m)\} \tag{1}$$

where $\{A_1, A_2, \ldots, A_m\}$ is the set of attributes in $K$ and $\mu_i$ is the membership of $O$ with attribute $A_i$ in $K$.

**Definition 2 (Fuzzy formal Concept)** Given $K = (G, M, I)$ and confidence threshold $T$, define:

- $A^* = \{m \in M | \forall g \in A : \mu(g, m) \geq T\}, A \subseteq G$
- $B^* = \{g \in G | \forall m \in B : \mu(g, m) \geq T\}, B \subseteq M$

A fuzzy formal concept is a pair $(A_f = \phi(A), B)$, where $A \subseteq G$, $B \subseteq M$, $A^* = B$, and $B^* = A$. Each object $g \in \phi(A)$ has a membership $\mu_g$ defined as

$$\mu_g = min_{m \in B} \mu(g, m) \tag{2}$$

**Definition 3 (Fuzzy formal Concept Similarity)** The similarity of a fuzzy formal concept $A_{f1} = (\phi(A_1), B_1)$ and $A_{f2} = (\phi(A_2), B_2)$ is defined as

$$E(A_{f1}, A_{f2}) = E(\phi(A_1), \phi(A_2)) \tag{3}$$

where $E(A, B) = \frac{|A \cap B|}{|A \cup B|}$, $A$ and $B$ are the fuzzy sets.

Let us give an example to illustrate these definitions. Table 1 lists 6 nouns, $\{help, support, time, part, place, medal\}$. these nouns appear as objects to certain verbs. The table lists their fuzzy membership as an object for 5 verbs $\{win, need, take, receive, play\}$. The table represents a fuzzy formal context. Using the definition above with a confidence threshold $T = 0.5$, a concept lattice is obtained as shown in Figure 1. The membership values for each object are obtained as defined in definition 2. The Figure 1 is the lattice of fuzzy formal concepts. The concept node $C_0$ is the root node with no attributes attached to it. The nodes $C_1$, $C_2$ and $C_3$ are the subconcepts of node $C_0$. A subconcept node inherits the attributes of its superconcept node. The node $C_1$ has the attribute set $\{win\}$ and the objects with fuzzy membership $\{support(0.7), medal(0.9)\}$. The node $C_4$ is a subconcept of node $C_1$ and $C_2$. It inherits the attributes of both the nodes and has an additional attribute $\{receive\}$. As we move down the hierarchy, number of attributes increases while the number of objects decreases. The node $C_7$ does not contain any of the objects and has all the 5 attributes.

| | win | need | take | receive | play |
|---|---|---|---|---|---|
| help | 0.0 | 0.8 | 0.2 | 0.3 | 0.0 |
| support | 0.7 | 0.9 | 0.0 | 0.7 | 0.0 |
| time | 0.0 | 0.8 | 0.65 | 0.2 | 0.0 |
| part | 0.0 | 0.1 | 0.8 | 0.0 | 0.7 |
| place | 0.1 | 0.3 | 0.8 | 0.2 | 0.0 |
| medal | 0.9 | 0.1 | 0.2 | 0.3 | 0.0 |

Table I
THE FORMAL CONTEXT FOR WORDS AND ATTRIBUTES FROM DATABASE

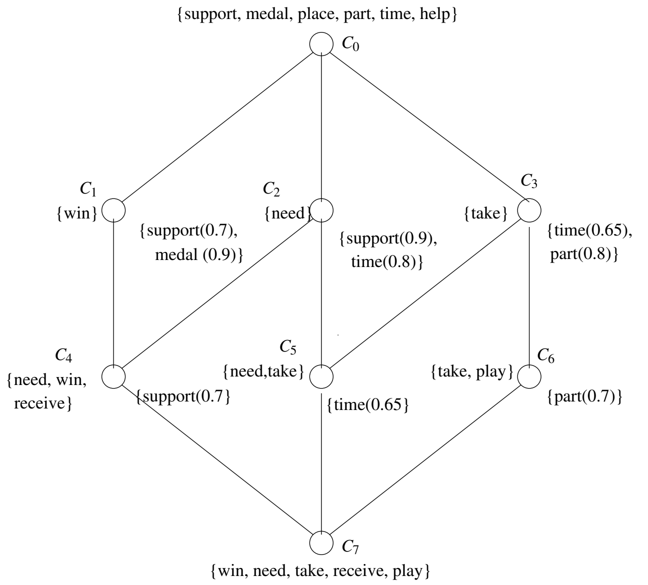

Figure 1. The lattice of formal concepts for the object-attribute pair in Table 1

It is desirable to give a label to each abstract concept. Cimiano [8] gives a label to each concept using the attribute concept [1]. However, an algorithm to label the concepts is

[1] assuming that the lattice is represented using reduced labeling [14]

necessary. In the next section, an algorithm for labeling the concept node with the concept summarizer has been described.

## III. Concept summarization Using Soft Least Upper Bound

It is useful to introduce 'meaningful' concept descriptors for the purpose of easier human readability. For instance, in Figure 1, the node C1 contains two objects $\{support, medal\}$ with their fuzzy memberships. For human readability, it will be good if the node can be labeled by a concept which summarizes 'support' and 'medal'. For this purpose, the concept summarization approach, presented by Andreasen [13] has been used. The summarization approach is strictly ontology based. Wordnet has been used as a general ontology. Certain terms need to be defined, before the assumptions and methodology can be discussed.

### A. Attribution of a concept

Attribution is a structure through which, compound concepts can be obtained from atomic concepts. Given atomic concepts $A$ and semantic relations $R$, the set of well-formed terms $L$ can be defined as:

$$L = \{A\} \cup \{x[r_1 : y_1, \ldots, r_n : y_n] | x \in A, r_i \in R, y_i \in L\} \quad (4)$$

for instance, with $R = \{TMP, LOC, \ldots\}$ and $A = \{town, old, \ldots\}$, we can have attributions like:

$$L = \{town, old, town[TMP : old], \ldots\}$$

### B. General Ontology

The general ontology $O = (L, \leq, R)$ encompasses

- A set of well-formed expressions $L$ as defined in subsection III-A.
- An inclusion relation "$\leq$" generalized from the taxonomy relation in $T$[2].
- A supplementary set of semantic relations $R$.

For $r \in R$, we have $x[r : y] \leq x$ and that $x[r : y]$ is in relation $r$ to $y$.

### C. Instantiated Ontology

Given a general ontology $O = (L, \leq, R)$ and a set of concepts $C$, the instantiated ontology $O_C = (L_C, \leq_C, R)$ is a restriction of $O$ to cover only the concepts in $C$:

$$L_C = C \cup \{x | y \in C, x \in L, y \leq x\} \quad (5)$$

$$\text{“} \leq_C \text{”} = \{(x, y) | x, y \in L_C, x \leq y\} \quad (6)$$

An assumption is made that any given object in a concept node is either a concept node, present in Wordnet or can be obtained from Wordnet by the use of an attribution relation. Thus it is possible to get an instantiated hierarchy from the Wordnet for a concept node. This instantiated hierarchy is used to derive an appropriate summary that grasps what is most characteristic about $C$. The problem is to find the least upper bound. Least upper bound(lub) is of interest when a common origin of two or more nodes is to be found. For example, 'cat' and 'dog' are connected though the concept 'mammal', while 'cat' and 'snake' are connected through 'animal'. As shown in [13], the fuzzified summary is noise tolerant, i.e. if there is a noisy object in the concept node, its effect on finding the least upper bound will be minimized.

### D. Soft least upper bound based approach

A sot definition of '*lub*' for a set of concepts $C'$ comprises of "upper boundedness" as well as "leastness" expressing respectively the concepts in $C'$ that are generalized and the degree to which a concept is least upper with regards to one or more concepts in $C'$.

"upper boundedness" can be expressed as the support:

$$\mu_{ub(C')}(x) = support(x, C') \quad (7)$$

where $support(x, C')$ for a given concept specifies the fraction of elements from the $C'$ covered

$$support(x, C') = \frac{|\{y | y \in C', y \leq x\}|}{|C'|} \quad (8)$$

Thus "upper boundedness" covers all generalizations of one or more concepts in $C'$ as full members.
"Leastness" can be defined as to express how close a concept is to a set of concepts:

$$\mu_{lu(C)}(x) = \{1 - \frac{dist(C', x)}{dist(C', Top)}\} \quad (9)$$

$$dist(C', y) = min_{x \in C'} dist(x, y) \quad (10)$$

where $dist(x, y)$ expresses the shortest upward path from $x$ to $y$. '$Top$' is the 'root' node in the hierarchy. "leastness" is thus defined with respect to the 'root' node in the tree.

Soft least upper bound can be defined as

$$\mu_{flub(C')}(x) = \mu_{lu(C)}(x) * \mu_{ub(C')}(x) \quad (11)$$

In our formalism, since the fuzzy representation of objects is being used, the definition of support has been modified as:

$$support(x, C') = \frac{|\{\sum_j \mu_j | c_j \in C', c_j \leq x\}|}{\sum_{i=1}^{N} \mu_i} \quad (12)$$

where $\{c_1, \ldots, c_N\}$ are the concepts in $C'$ with the membership values $\{\mu_1, \ldots, \mu_N\}$. The generalized concept with the highest membership value is used as the resultant summarizer.

As will be described in section V, experiments were performed on British National corpus[3] to extract the verb/subject and verb/object syntactic dependencies varying the number of verbs, $N$ and the confidence threshold $T_s$. 5

[2] The presence of a taxonomy $T$ over the set of atomic concepts $A$ has been assumed

[3] http://www.natcorp.ox.ac.uk

object attribute pairs are extracted, obtained from one such experiment with $N = 30$ and $T_s = 0.03$. The syntactic relation 'dobj (direct object)' was used for generating the hierarchy. We enumerate the objects and attributes of nodes 6, 20, 19, 59 and 21, represented in Figure 2

```
objects[06]:    career, proceeding, process,
                program,work
attributes[06]: begin
object[20]:     business work
attributes[20]: do start
object[19]:     business course
attributes[19]: run start
object[21]:     business
attributes[21]: do run start
object[59]:     work
attributes[59]: begin do start
```

Figure 2 shows only that part of the generated concept hierarchy, which contains these 5 pairs. Node 0 is the root node. The concept summarization is then used for each generated hierarchy. The results for the soft least upper bound based approach are as follows:

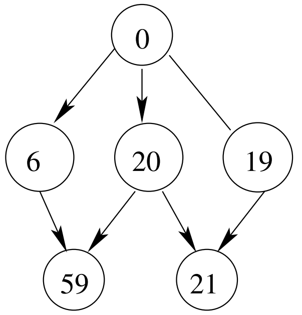

Figure 2. A part of the concept hierarchy, generated in the experiments

- Node 6: The cluster is $\{career, proceeding, process, program, work\}$. The summarizers obtained are:

| | |
|---|---|
| occupation | 0.17142858 |
| activity | 0.5 |
| act | 0.5333333 |
| event | 0.4 |
| psychological feature | 0.3333333 |

- Node 20: The cluster is $\{business, work\}$. The summarizers obtained are:

| | |
|---|---|
| enterprise | 0.333 |
| organization | 0.3333333 |
| abstraction | 0.16666669 |
| activity | 0.41666666 |
| act | 0.3333333 |

- Node 19: The cluster is $\{business, course\}$. The summarizers obtained are:

| | |
|---|---|
| enterprise | 0.41666666 |
| organization | 0.3333333 |
| education | 0.3283 |
| activity | 0.35714287 |
| act | 0.28571427 |

For each non-leaf node, a fuzzy set of summarizers is obtained using the modified definition of support (Equation 12). The summarizer with maximum fuzzy value is selected to label the corresponding node. Act, Activity and Enterprise are obtained as the summarizers for the node 6, 20 and 19 respectively. The concept hierarchy in Figure 2 can now be represented as Figure 3 using the summarizers for node 6, 20 and 19[4]. All the leaf nodes are not shown in the figure.

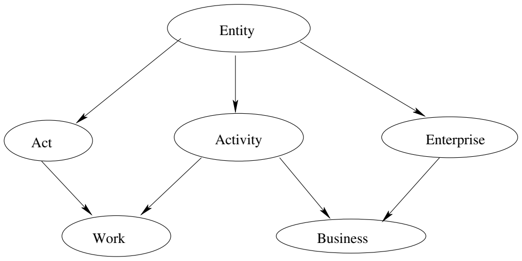

Figure 3. The hierarchy shown in Figure 2 after labeling the concepts using concept summarization

## IV. Modified Fuzzy Relational Ontological Model in Information Search

In this section, we describe a modified fuzzy relational ontological model for information search. The document and query representation are described first. These representations have been borrowed from Pereira [4]. Let $I$ be the fuzzy formal context of the words and attributes, represented as

$$I = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix} \tag{13}$$

where $1 \leq i \leq n$, $n$ is the number of words; $1 \leq j \leq m$, $m$ is the number of attributes(categories) and $r_{ij} \in [0,1]$ is the relevance degree between word $O_i$ and attribute $A_j$.

### A. Query Representation

Let $Q = \{k_i, c_j\}$ be the set of words and categories in the user query. The query $q$ is represented

[4] In wordnet, Entity is the root node and therefore, node 0 has been labeled with 'Entity'

by vectors $x = [x_1, x_2, \ldots, x_i, \ldots, x_n], 1 \leq i \leq n$, and $y = [y_1, y_2, \ldots, y_j, \ldots, y_m], 1 \leq j \leq m$, such that:

$$x_i = \begin{bmatrix} 1 & if O_i \in Q \\ 0 & otherwise \end{bmatrix}, y_j = \begin{bmatrix} 1 & if A_j \in Q \\ 0 & otherwise \end{bmatrix} \tag{14}$$

### B. Document Representation

Let $D$ be the collection of documents, $D = \{d_1, d_2, \ldots, d_{doc}, \ldots, d_u\}$. Documents can be represented by matrix $T_k$ as follows:

$$T_k = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{u1} & \alpha_{u2} & \ldots & \alpha_{un} \end{bmatrix} \tag{15}$$

where $\alpha_{doc,i} \in [0,1]$ is the relevance degree between the document $d_{doc}$ and the word $O_i, 1 \leq doc \leq u$, $1 \leq i \leq n$; and by a $T_c$ matrix

$$T_c = \begin{bmatrix} \beta_{11} & \beta_{12} & \ldots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \ldots & \beta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{u1} & \beta_{u2} & \ldots & \beta_{um} \end{bmatrix} \tag{16}$$

where $\beta_{doc,j} \in [0,1]$ is the relevance degree between the document $d_{doc}$ and the attribute $A_j$, $k_i, 1 \leq doc \leq u$, $1 \leq j \leq m$.

### C. Information retrieval with fuzzy formal concepts

For the information retrieval, propositions from the fuzzy formal concepts are extracted. A proposition can be represented as "$x$ is $V$", where $x$ is a variable and $V$ is a value. In fuzzy logic, the proposition can be represented as a fuzzy set $U$, which implies "$x$ is $U$". The widely used proposition in fuzzy logic is the "IF-THEN" proposition. Assume that a formal concept $C_i$ has a set of objects $\{O_1, O_2, \ldots, O_n\}$. An "IF-THEN" proposition can be constructed as follows: "IF $x$ is $O_1$ or $x$ is $O_2$...or $x$ is $O_n$, then $x$ belongs to $C_i$". However, objects can be shared between formal concepts and some objects in $C_i$ may belong to objects in $C_j$. Therefore, the "THEN" part of the proposition needs to be fuzzified as well. It needs to contain all the formal concepts alongwith their fuzzy membership values. The membership value of $C_j$ in the THEN part of the proposition should be the fuzzy formal concept similarity, as defined in Definition 3.

The IF part of the proposition can be represented as a fuzzy set $\phi(O_1) \cup \phi(O_2) \cup \ldots \cup \phi(O_n)$. Thus, for $N$ formal concepts, $N$ propositions will be obtained. Any such proposition $R_i, 1 \leq i \leq N$ can be written using max-min composition [15].

The fuzzy representation of IF part is:

$$\phi = \begin{bmatrix} \mu_1 & \mu_2 & \ldots & \mu_m \end{bmatrix}' \tag{17}$$

where $\mu_i, 1 \leq i \leq m$ is the fuzzy membership of attribute $A_i$ obtained using $\phi(O_1) \cup \phi(O_2) \cup \ldots \cup \phi(O_n)$.

The fuzzy representation of THEN part is:

$$S = \begin{bmatrix} a_{i1} & a_{i2} & \ldots & a_{iN} \end{bmatrix} \tag{18}$$

where $a_{ij} = E(\phi(O_i), \phi(O_j)), 1 \leq j \leq N$[5]

$$R_i = \phi \circ S \tag{19}$$

If $I$ is the fuzzy formal context, the query vectors $x$ and $y$ are processed as follows:

A fuzzy set $G_c$ is obtained from $x$ and $I$ as follows:

$$G_c = x \circ I \tag{20}$$

The attribute (category) vector is represented as

$$G = G_c \cup y \tag{21}$$

The extracted propositions are used to get the membership of formal concepts as:

$$\delta_G = \bigcup_{i=1}^{N} G \circ R_i \tag{22}$$

The document vector can be processed to obtain as

$$T = T_k \circ I \cup T_c \tag{23}$$

and the extracted propositions can be used to get the membership of formal concepts for all the documents as:

$$\delta_T = \bigcup_{i=1}^{N} T \circ R_i \tag{24}$$

The relevance degree between the document and the query can be computed as

$$V_{DQ} = \delta_T \circ \delta'_G \tag{25}$$

Let us give an example to explain the procedure:

| | Train | Describe | Evaluate |
|---|---|---|---|
| Neural Network | 0.7 | 0.4 | 0.9 |
| Information Retrieval | 0.3 | 0.9 | 0.2 |
| Learning algorithm | 0.1 | 0.8 | 0.7 |

Table II
THE FORMAL CONTEXT FOR THREE (KEY)WORDS AND ATTRIBUTES

The Table 2 gives the fuzzy formal context of three words, alongwith the attributes. These attributes are to be extracted from the abstracts or the complete document. Now, as discussed in section II, the fuzzy formal concept lattice is built as shown in Figure 4 (Confidence threshold $T = 0.5$ is used for creating the lattice). The similarities between the formal concepts are shown. The process for proposition extraction is: For the

[5] $E(A,B)$ is defined in Definition 3

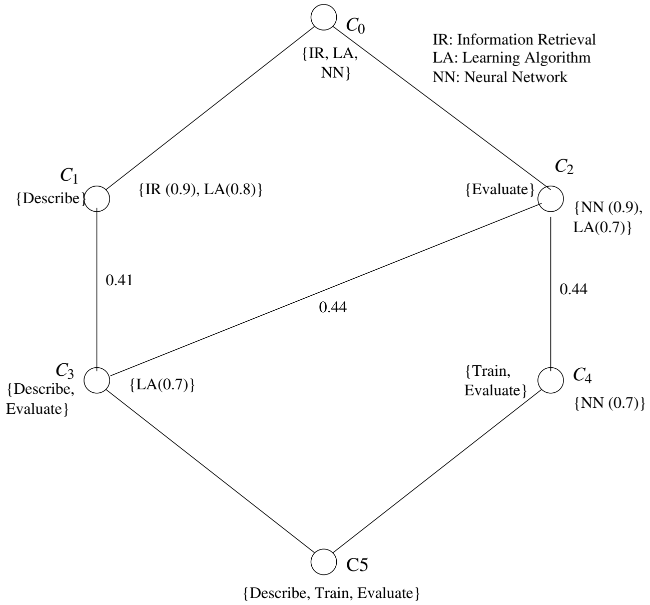

Figure 4. The lattice for keywords-attribute pair in Table 2

concept node $C_1$, the IF part has the fuzzy representation (Equation 17) $\phi = \begin{bmatrix} 0.0 & 0.9 & 0.7 \end{bmatrix}'$, obtained using $\phi(InformationRetrieval) \cup \phi(LearningAlgorithm)$. The THEN part of the proposition is modeled (Equation 18) as

$$S = \begin{bmatrix} 1.0 & 0.0 & 0.41 & 0.0 \end{bmatrix}$$

The proposition can be represented (Equation 19) as

$$R_1 = \begin{bmatrix} 0.0 \\ 0.9 \\ 0.7 \end{bmatrix} \circ \begin{bmatrix} 1.0 & 0.0 & 0.41 & 0.0 \end{bmatrix} \quad (26)$$

$$= \begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.0 \\ 0.9 & 0.0 & 0.37 & 0.0 \\ 0.7 & 0.0 & 0.29 & 0.0 \end{bmatrix} \quad (27)$$

Suppose 3 documents are given with their matrices as follows:

$$T_k = \begin{bmatrix} 0.3 & 0.8 & 0.9 \\ 0.7 & 0.2 & 0.6 \\ 0.8 & 0.7 & 0.1 \end{bmatrix}$$

$$T_c = \begin{bmatrix} 0.4 & 0.7 & 0.2 \\ 0.8 & 0.9 & 0.1 \\ 0.7 & 0.8 & 0.7 \end{bmatrix} \quad (28)$$

The document vector is (Equation 23)

$$\begin{aligned} T &= T_k \circ I \cup T_c \\ &= \begin{bmatrix} 0.3 & 0.8 & 0.7 \\ 0.7 & 0.6 & 0.7 \\ 0.7 & 0.7 & 0.8 \end{bmatrix} \circ \begin{bmatrix} 0.4 & 0.7 & 0.2 \\ 0.8 & 0.9 & 0.1 \\ 0.7 & 0.8 & 0.7 \end{bmatrix} \\ &= \begin{bmatrix} 0.4 & 0.8 & 0.7 \\ 0.8 & 0.6 & 0.7 \\ 0.7 & 0.8 & 0.8 \end{bmatrix} \quad (29) \end{aligned}$$

The membership of formal concepts using the proposition $R_1$ can be obtained as

$$\begin{aligned} T \circ R_1 &= \begin{bmatrix} 0.4 & 0.8 & 0.7 \\ 0.8 & 0.6 & 0.7 \\ 0.7 & 0.8 & 0.8 \end{bmatrix} \circ \begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.0 \\ 0.9 & 0.0 & 0.37 & 0.0 \\ 0.7 & 0.0 & 0.29 & 0.0 \end{bmatrix} \\ &= \begin{bmatrix} 0.8 & 0.0 & 0.296 & 0.0 \\ 0.7 & 0.0 & 0.222 & 0.0 \\ 0.8 & 0.0 & 0.296 & 0.0 \end{bmatrix} \quad (30) \end{aligned}$$

The matrix $\delta_T$ can be obtained by taking the union over all the propositions (Equation 24). The computation of the query matrix $\delta_G$ will follows the same process. The relevance between the document and query is computed using Equation 25.

## V. Results

First, the experiment on automatically building the domain ontology has been described. As mentioned in the introduction, we make use of the syntactic dependencies between the verbs and nouns to describe the context attributes of the terms, we are interested in. These dependencies have been extracted automatically using the Stanford parser [16], which is a trainable, statistical state-of the art parser. The parser gives the dependency parse in addition to the parse tree. The syntactic dependencies are extracted using the dependency parse. Finally, the verbs as well as the head of objects are lemmatized by using the lemma information, which is extracted from the corpus and stored in a finite state transducer (FST). However, the results from text processing can not be directly used in the formal concept analysis, because not all derived dependencies are 'interesting'. To deal with this problem, a conditional information measure [17] has been used to weight the object/attribute pairs. The confidence threshold, as defined in Section II processes only those verb/argument relations for which this measure is above this threshold. The conditional information measure can be represented as

$$Conditional(n, v_{arg}) = P(n|v_{arg}) = \frac{f(n, v_{arg})}{f(v_{arg})} \quad (31)$$

where $f(n, v_{arg})$ is the total number of occurrences of a term $n$ as argument $arg$ of a verb $v$, $f(v_{arg})$ is the number of occurrences of verb $v$ with such an argument.

### A. Algorithm for Text to Ontology Representation

The algorithm for automatically learning the ontology from text documents is as follows:

1) The text documents are pre-processed and parsed.
2) The syntactic dependencies are extracted from the parse tree.
3) The syntactic dependencies are lemmatized using a finite state transducer.
4) The syntactic dependencies are weighted using the conditional information measure.

|  | Tourism | Finance |
|---|---|---|
| No. Concepts | 293 | 1223 |
| No. Leaves | 236 | 861 |
| Avg. Depth | 3.99 | 4.57 |
| Max. Depth | 6 | 13 |
| Max. children | 21 | 33 |
| Avg. Children | 5.26 | 3.5 |

Table III
ONTOLOGY STATISTICS

5) Only those dependencies, for which the information measure $> T_s$ and one of the $N$ most frequent verbs appear as attribute, are used.
6) The fuzzy formal context is obtained.
7) The fuzzy formal concepts are obtained using the definitions in section II.
8) The lattice $C'$ is represented using reduced labeling[6].
9) Concept summarization is applied over all non-leaf nodes, as per the algorithm given in subsection V-B.

### *B. Algorithm for Labeling the concept Nodes using Concept Summarization*

Let $N$ be the set of terms, we are interested in. The fuzzy formal concept analysis is applied to get the concept lattice. Let $T = N/L$, where $L$ is the set of leaf nodes, as obtained in the lattice $C'$. The algorithm for labeling the non-leaf nodes is as follows (We start with the root node $V$):

1) Let $O = \{O_1, \ldots, O_i, \ldots, O_J\}$ be the fuzzy set of the objects at node $V$.
2) Apply the soft least upper bound approach (Equations 9-12) over set $O$ to get the $K$ summarizers $M_V = \{m_1, \ldots, m_K\}$ and their fuzzy membership values $F_V = \{\mu_1, \ldots, \mu_K\}$ (Only the membership values above a threshold α are to be used. $\mu_i > \alpha, 1 \leq i \leq K$)
3) The summarizer concepts is

$$L_V = \begin{bmatrix} max_j(\mu_j), 1 \leq j \leq K & if \mu_j \notin T, 1 \leq j \leq K \\ max_j(\mu_j), 1 \leq j \leq K, \mu_j \in T & otherwise \end{bmatrix} \tag{32}$$

4) $T = T/L_V$.
5) For all the subconcepts of node $V$: apply steps 1 to 5.

### *C. Ontology Evaluation*

Increased use of domain ontologies requires well-established methods to evaluate them. For the evaluation purposes, two domain ontologies have been taken: tourist and finance. These ontologies are modeled by experienced ontology engineers. Table 3 summarizes some facts about these ontologies. A general purpose corpus, the British National Corpus was used for the experiments. The corpus data is in xml format. The xml tree was parsed using Java DOM parser and the lemma and the sentences were stored in separate files. Lemmas obtained were used to make a finite state transducer to be used for lemmatization step. After the lemmatization, $N$ most frequent verbs were selected and used the fuzzy formal concept analysis is applied over the nouns with a confidence threshold $T_s$. At each concept node, the objects are summarized using the fuzzy based least upper bound approach. For ontology evaluation, the Semantic Cotopy, proposed in [8] has been used.

The Common Semantic Cotopy of a concept $c_i \in O_1$ with respect to $O_2$ is defined as:

$$SC(c_i, O_1, O_2) := \{c_j \in C_1 \cap C_2 | c_j <_{C_1} c_i \wedge c_i <_{C_1} c_j\} \tag{33}$$

where $C_1$ is the set of all concepts in ontology $O_1$ and $C_2$ is the set of all concepts in ontology $O_2$. The taxonomic overlap $TO(O_1, O_2)$ between the two ontologies is calculated by averaging over all the taxonomic overlaps of the concepts in $C_1$. As proposed in [8], we do not calculate the semantic cotopy for concepts which are common in both the ontologies. The Taxonomic overlap between two ontologies is calculated as follows:

$$TO(O_1, O_2) = \frac{1}{|C_1/C_2|} \sum_{c \in C_1/C_2} max_{c' \in C_2 \cup \{root\}}$$

$$\frac{|SC(c, O_1, O2) \cap SC(c', O_2, O_1)|}{|SC(c, O_1, O2) \cup SC(c', O_2, O_1)|} \tag{34}$$

The precision, recall and F-Measure are calculated as follows:

$$P(O_1, O_2) = TO(O_1, O_2) \tag{35}$$

$$R(O_1, O_2) = TO(O_2, O_1) \tag{36}$$

$$F(O_1, O_2) = \frac{2.P(O_1, O_2).R(O_1, O_2)}{P(O_1, O_2) + R(O_1, O_2)} \tag{37}$$

For our experiments using the two ontologies, the results obtained are shown in Table 4 and 5. The two approaches using formal concept analysis (FCA) and fuzzy formal concept analysis with soft least upper bound (LubFCA) have been compared for both the ontologies. The number of verbs $N$ has been fixed to 250 and all the dependency relations are used. The results have been described for two different thresholds with $T_s = 0.01$ and $T_s = 0.005$.

|  | Precision | Recall | F-Measure |
|---|---|---|---|
| FCA($T_s = 0.01$) | 32.68% | 50.91% | 39.81% |
| LubFCA($T_s = 0.01$) | 30.14% | 60.42% | 40.21% |
| FCA($T_s = 0.005$) | 34.21% | 48.92% | 40.26% |
| LubFCA($T_s = 0.005$) | 31.34% | 58.22% | 40.74% |

Table IV
COMPARISON OF RESULTS FOR TOURISM DOMAIN FOR VARIOUS VALUES OF $T_s$

As is clear from the results, LubFCA approach shows a decrease in precision, however, the recall is increased. The precision obtained was less due to the following reasons:

[6] Reduced labeling as defined in [14] means that objects are in extension of the most specific concept and attributes conversely in the intension of the most general one.

| | Precision | Recall | F-Measure |
|---|---|---|---|
| FCA($T_s = 0.01$) | 27.62% | 36.01% | 31.26% |
| LubFCA($T_s = 0.01$) | 25.14% | 42.21% | 31.51% |
| FCA($T_s = 0.005$) | 29.21% | 34.92% | 31.35% |
| LubFCA($T_s = 0.005$) | 27.34% | 41.22% | 32.88% |

Table V
COMPARISON OF RESULTS FOR FINANCE DOMAIN FOR VARIOUS VALUES OF $T_s$

- Wordnet was used for the concept summarization. However, it was not clear as to how the word sense disambiguation algorithm can be used to pick up the appropriate sense for a given term. For each word, the first sense given by the wordnet was used.
- The adopted measure for precision (Equation 35) uses the concepts which are common in both the ontologies. In LubFCA approach, the number of these concepts decreases and it was one of the reasons for decreased precision values.

As a future work, an appropriate setting for using the correct sense of Wordnet needs to be used and different measures for ontology evaluation need to be tested.

## VI. CONCLUSIONS

The paper presents a novel approach for labeling the concept nodes of automatically built domain ontologies. The fuzzy least upper bound formulation has been modified (Equation 12) to take into account the fuzzy membership of objects in a cluster. The complete algorithm has been proposed in the section V. The formalization is promising since, in addition to learning concept hierarchies using formal concept analysis, it presents a framework for giving a meaningful identifier for the concepts, which is a summarizer over all the objects included in the concept. As shown by the experiments, the approach improves the recall and F-Measure for the two domain ontologies considered. An application of concept hierarchies for information retrieval is presented, which modifies the previously proposed fuzzy relational ontological model using fuzzy formal concepts. The mathematical formulation for the information retireval application is presented in the paper and it needs to be tested over a set of documents with respect to user queries.

Future work will include the experiments over the modified fuzzy relational model for information retrieval. The applications of concept hierarchies for question answering will also be investigated.

## REFERENCES

[1] Y.-J. Horng, S.-M. Chen, and C.-H. Lee, "Automatically constructing multi-relationship fuzzy concept networks for document retrieval," *Applied Artificial Intelligence*, vol. 17, no. 4, pp. 303–328, 2003.

[2] Y. Ogawa, T. Morita, and K. Kobayashi, "A fuzzy document retrieval system using the keyword connection matrix and a learning method," *Fuzzy Sets Syst.*, vol. 39, no. 2, pp. 163–179, 1991.

[3] S. K. Pal, V. Talwar, P. Mitra, and S. Members, "Web mining in soft computing framework: Relevance, state of the art and future directions," *IEEE Transactions on Neural Networks*, vol. 13, pp. 1163–1177, 2002.

[4] R. Pereira, I. Ricarte, and F. Gomide, *Elie Sanchez. (Org.). Fuzzy Logic and The Semantic Web*. Amsterdan: Elsevier, 2006, ch. Fuzzy relational ontological model in information search systems, pp. 395–412.

[5] A. Hotho, S. Staab, and G. Stumme, "Ontologies improve text document clustering," in *ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2003, p. 541.

[6] S. Bloehdorn and A. Hotho, "Text classification by boosting weak learners based on terms and concepts," in *ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2004, pp. 331–334.

[7] E. M. Voorhees, "Query expansion using lexical-semantic relations," in *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: Springer-Verlag New York, Inc., 1994, pp. 61–69.

[8] P. Cimiano, A. Hotho, and S. Staab, "Learning concept hierarchies from text corpora using formal concept analysis," *Journal of Artificial Intelligence research*, vol. 24, pp. 305–339, 2005.

[9] Z. Harris, *Mathematical Structures of Language*. Wiley.

[10] R. Navigli and P. Velardi, "Learning domain ontologies from document warehouses and dedicated web sites," *Comput. Linguist.*, vol. 30, no. 2, pp. 151–179, 2004.

[11] Q. T. Tho and T. H. Cao, "Automatic fuzzy ontology generation for semantic web," *IEEE Trans. on Knowl. and Data Eng.*, vol. 18, no. 6, pp. 842–856, 2006, senior Member-Hui,, Siu Cheung and Senior Member-Fong,, A. C. M.

[12] A. Zouaq and R. Nkambou, "Building domain ontologies from text for educational purposes," *IEEE Trans. Learn. Technol.*, vol. 1, no. 1, pp. 49–62, 2008.

[13] T. Andreasen and H. Bulskov, "Conceptual querying through ontologies," *Fuzzy Sets and Systems*, vol. 160, no. 15, pp. 2159 – 2172, 2009.

[14] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1999, translator-Franzke,, C.

[15] G. J. Klir and B. Yuan, *Fuzzy sets and fuzzy logic*. Upper Saddle River: Prentice-Hall, 1995.

[16] B. MacCartney, "The stanford parser version 1.6," 2008.

[17] P. Cimiano, "Ontology-driven discourse analysis in genie," in *NLDB*, 2003, pp. 77–90.