

FOGA: A Fuzzy Ontology Generation Framework for Scholarly Semantic Web

Thanh Tho Quan¹, Siu Cheung Hui¹ and Tru Hoang Cao²

¹School of Computer Engineering, Nanyang Technological University,
Singapore

{PA0218164B, asschui}@ntu.edu.sg

²Faculty of Information Technology, Hochiminh City University of Technology,
Hochiminh City, Vietnam

tru@dit.hcmut.edu.vn

Abstract. This paper proposes the FOGA (Fuzzy Ontology Generation framework) for automatic generation of fuzzy ontology on uncertainty information. The FOGA framework comprises the following components: Fuzzy Formal Concept Analysis, Fuzzy Conceptual Clustering and Fuzzy Ontology Generation. First, Fuzzy Formal Concept Analysis incorporates fuzzy logic into Formal Concept Analysis (FCA) to form a fuzzy concept lattice. Fuzzy Conceptual Clustering then constructs the concept hierarchy from the fuzzy concept lattice. Finally, Fuzzy Ontology Generation generates the fuzzy ontology from the concept hierarchy. In this paper, we will also discuss the application of the FOGA framework to generate a scholarly ontology for the Scholarly Semantic Web from a citation database. The performance of the proposed FOGA framework is given based on the scholarly ontology generated.

Keywords: Formal Concept Analysis, Fuzzy Logic, Conceptual Clustering, Ontology Generation

1 Introduction

Ontology is a conceptualization of a domain into a human understandable, but machine-readable format consisting of entities, attributes, relationships and axioms [1]. Ontology uses classes to represent concepts. Ontology also supports taxonomy and non-taxonomy relations between classes. However, the conceptual formalism supported by typical ontology may not be sufficient to represent uncertainty information that is commonly found in many application domains. For example, keywords extracted from scientific publications can be used to infer the corresponding research areas, however, it is inappropriate to treat all keywords equally as some keywords may be more significant than others. In addition, it is sometimes difficult to judge whether a document belongs completely to a research area or not.

To tackle this type of problems, one possible solution is to incorporate fuzzy logic into ontology to handle uncertainty data. Traditionally, fuzzy ontology is

generated and used in text retrieval [2], in which membership values are used to evaluate the similarities between concepts on a concept hierarchy. However, the pure manual generation of fuzzy ontology from a predefined concept hierarchy is a difficult and tedious task that is often required expert interpretation. As such, automatic generation of concept hierarchy and fuzzy ontology from uncertainty data of a domain is needed.

In this paper, we propose a framework known as FOGA (Fuzzy Ontology Generation frAMework) that can automatically generate a fuzzy ontology on uncertainty data. As compared with existing fuzzy ontology generation techniques, FOGA can automatically construct a hierarchy structure of ontology classes. In addition, this paper also discusses the use of FOGA to generate scholarly ontology for the Scholarly Semantic Web from an experimental citation database. Here, the taxonomy relations on ontology classes can be generated automatically as compared with the manual method used in other semantic scholarly systems such as ESKIMO [4]. However, FOGA still requires some minimal human interpretations to help add meaningful labels on initial class names, attributes and its relations.

The rest of this paper is organized as follows. Section 2 discusses the related works on ontology generation and FCA. Section 3 describes the FOGA framework. Section 4 discusses Fuzzy Formal Concept Analysis (FFCA). Section 5 discusses conceptual clustering based on FFCA. Fuzzy ontology generation is given in Section 6. The use of FOGA for generating a scholarly ontology for Scholarly Semantic Web using an experimental citation database is discussed in Section 7. Section 8 gives the performance evaluation. Finally, Section 9 concludes the paper.

2 Related Works

There are many techniques such as Natural Language Processing (NLP) combined with association rules [5], statistical model [6], and clustering [7] that have been applied to generate ontology from a concept hierarchy automatically or semi-automatically. Among them, clustering is one of the most effective techniques for ontology learning. Moreover, conceptual clustering techniques such as COBWEB are powerful clustering techniques that can be used for the generation of concept representations and relationships for ontology [8].

FCA is a formal technique for data analysis and knowledge presentation. Recently, FCA is an effective technique for conceptual clustering as discussed in Section 1. However, as most concept lattices are quite complicated in terms of the number of concepts generated, it is necessary to simplify the lattice generated. In Iceberg concept lattice [9], association rules are used to cluster concepts on the lattice, and conceptual scaling [10] is then used to generate the concept hierarchy.

As discussed in Section 1, fuzzy logic can be incorporated into FCA to handle uncertainty information for conceptual clustering and concept hierarchy generation. Pollandt [11] have proposed the L-Fuzzy context as an attempt to combine

fuzzy logic with FCA. The L-Fuzzy context uses linguistic variables, which are linguistic terms associated with fuzzy sets, to represent uncertainty in the context. However, human interpretation is required to define the linguistic variables. Moreover, the fuzzy concept lattice generated from the L-fuzzy context usually causes a combinatorial explosion of concepts as compared to the traditional concept lattice.

In this paper, we propose a new technique that incorporates fuzzy logic and FCA as Fuzzy Formal Concept Analysis (FFCA), in which the uncertainty information is directly represented by a real number of membership value in the range of $[0,1]$. As such, linguistic variables are no longer needed. In comparison with the fuzzy concept lattice generated from the L-fuzzy context, the fuzzy concept lattice generated using FFCA will be simpler in terms of the number of formal concepts, and it also supports a formal mechanism for calculating concept similarities. Therefore, the proposed FFCA's fuzzy concept lattice is a suitable representation for conceptual clustering.

3 The FOGA Framework

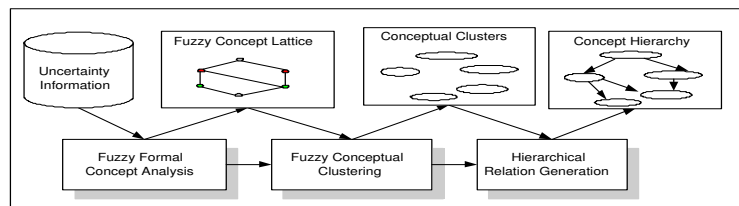


Fig. 1. The proposed approach for automatic generation of concept hierarchy

Figure 1 shows the proposed FOGA (Fuzzy Ontology Generation frAme-work), which consists of the following components:

- *Fuzzy Formal Concept Analysis* – It constructs a *fuzzy formal context* from a database containing uncertainty data. In addition, it also generates *fuzzy formal concepts* from the *fuzzy formal context* and organizes the generated concepts as a *fuzzy concept lattice*.
- *Fuzzy Conceptual Clustering* – It clusters concepts on the *fuzzy concept lattice* and generates *conceptual clusters*. The clustering process is performed based on fuzzy information incorporated into the lattice using fuzzy logic.
- *Hierarchical Relation Generation* – It generates hierarchical relations between conceptual clusters to construct a *concept hierarchy*.

4 Fuzzy Formal Concept Analysis

In FOGA, we propose the Fuzzy Formal Concept Analysis, which incorporates fuzzy logic into Formal Concept Analysis, to represent vague information.

Definition 1. A fuzzy formal context is a triple $K = (G, M, I = \varphi(G \times M))$ 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 $[0, 1]$.

A fuzzy formal context can also be represented as a cross-table as shown in Table 1(a). The context has three objects representing three documents, namely $D1$, $D2$ and $D3$. In addition, it also has three attributes, "Data Mining" (D), "Clustering" (C) and "Fuzzy Logic" (F) representing three research topics. The relationship between an object and an attribute is represented by a membership value between 0 and 1.

Table 1(a). A cross-table of a fuzzy formal context.

	D	C	F
D1	0.8	0.12	0.61
D2	0.9	0.85	0.13
D3	0.1	0.14	0.87

Table 1(b). Fuzzy formal context in Table 1(a) with $T = 0.5$.

	D	C	F
D1	0.8	-	0.61
D2	0.9	0.85	-
D3	-	-	0.87

A confidence threshold T can be set to eliminate relations that have low membership values. Table 1(b) shows the cross-table of the fuzzy formal context given in Table 1(a) with $T = 0.5$.

Generally, we can consider the attributes of a formal concept as the description of the concept. Thus, the relationships between the object and the concept should be the intersection of the relationships between the objects and the attributes of the concept. Since each relationship between the object and an attribute is represented as a membership value in fuzzy formal context, then the intersection of these membership values should be the minimum of these membership values, according to fuzzy theory [12].

Definition 2. Given a fuzzy formal context $K=(G, M, I)$ and a confidence threshold T , we define $A^* = \{m \in M | \forall g \in A: \mu(g, m) \geq T\}$ for $A \subseteq G$ and $B^* = \{g \in G | \forall m \in B: \mu(g, m) \geq T\}$ for $B \subseteq M$. A fuzzy formal concept (or fuzzy concept) of a fuzzy formal context (G, M, I) with a confidence threshold T is a pair $(A_f = \varphi(A), B)$ where $A \subseteq G$, $B \subseteq M$, $A^* = B$ and $B^* = A$. Each object $g \in \varphi(A)$ has a membership μ_g defined as

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

where $\mu(g, m)$ is the membership value between object g and attribute m , which is defined in I . Note that if $B = \{\}$ then $\mu_g = 1$ for every g .

Definition 3. Let (A_1, B_1) and (A_2, B_2) be two fuzzy concepts of a fuzzy formal context (G, M, I) . $(\varphi(A_1), B_1)$ is the subconcept of $(\varphi(A_2), B_2)$, denoted as $(\varphi(A_1), B_1) \leq (\varphi(A_2), B_2)$, if and only if $\varphi(A_1) \subseteq \varphi(A_2) (\Leftrightarrow B_2 \subseteq B_1)$. Equivalently, (A_2, B_2) is the superconcept of (A_1, B_1) .

Definition 4. A fuzzy concept lattice of a fuzzy formal context K with a confidence threshold T is a set $F(K)$ of all fuzzy concepts of K with the partial order \leq with the confidence threshold T .

Definition 5. The similarity of a fuzzy formal concept $K_1 = (\varphi(A_1), B_1)$ and its subconcept $K_2 = (\varphi(A_2), B_2)$ is defined as $E(K_1, K_2) = \frac{|\varphi(A_1) \cap \varphi(A_2)|}{|\varphi(A_1) \cup \varphi(A_2)|}$.

Figure 2 gives the traditional concept lattice generated from Table 1(a). Figure 3 gives the fuzzy concept lattice generated from the fuzzy formal context given in Table 1(b). As shown from the figures, the fuzzy concept lattice can provide additional information, such as membership values of objects in each fuzzy formal concept and similarities of fuzzy formal concepts, that are important for the construction of concept hierarchy.

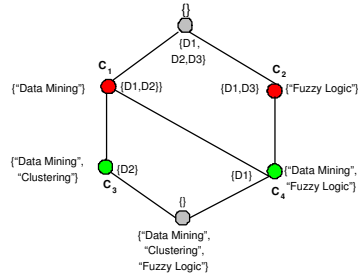


Fig. 2. A concept lattice generated from traditional FCA.

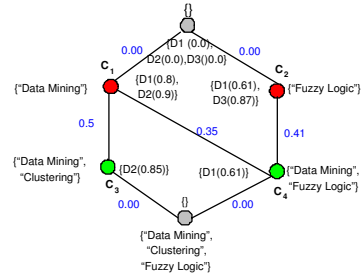


Fig. 3. A fuzzy concept lattice generated from FFCA.

5 Fuzzy Conceptual Clustering

As in traditional concept lattice, the formal concepts are generated mathematically, objects that have small differences in terms of attribute values are classified into distinct formal concepts. At a higher level, such objects should belong to the same concept when they are interpreted by human. Based on this observation, we propose to cluster formal concepts into conceptual clusters using fuzzy conceptual clustering. The conceptual clusters generated have the following properties:

- Conceptual clusters have hierarchical relationships that can be derived from fuzzy formal concepts on the fuzzy concept lattice. That is, a concept represented by a conceptual cluster can be a subconcept or superconcept of other concepts represented by other conceptual clusters.

- A formal concept must belong to at least one conceptual cluster, but it can also belong to more than one conceptual cluster. This property is derived from the characteristic of concepts that an object can belong to more than one concept. For example, a scientific document can belong to more than one research area.

Conceptual clusters are generated based on the premise that if a formal concept A belongs to a conceptual cluster R , then its subconcept B also belongs to R if B is similar to A . We can use a *similarity confidence threshold* T_s to determine whether two concepts are similar or not.

Definition 6. A conceptual cluster of a concept lattice K with a similarity confidence threshold T_s is a sublattice S_K of K which has the following properties:

1. S_K has a supremum concept C_S that is not similar to any of its superconcepts.
2. Any concept $C \neq C_S$ in S_K must have at least one superconcept $C' \in S_K$ such that $E(C, C') > T_s$.

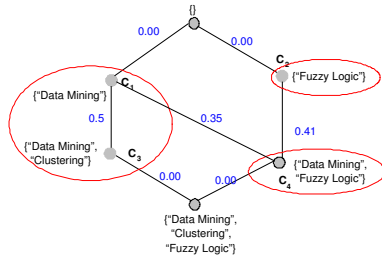


Fig. 4. Conceptual clusters.

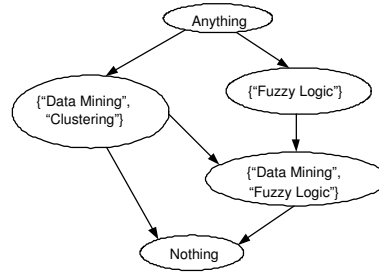


Fig. 5. Concept hierarchy.

Figure 4 shows the conceptual clusters that are generated from the concept lattice given in Figure 3 with the similarity confidence threshold $T_s = 0.5$. Figure 5 shows the corresponding concept hierarchy, in which each concept is represented by a set of attributes of objects from the corresponding conceptual cluster.

Figure 6 gives the algorithm that generates conceptual clusters from a concept C_S which is called the *starting concept* on a fuzzy concept lattice $F(K)$. To generate all conceptual clusters of $F(K)$, we choose C_S as the supremum of $F(K)$, or $C_S = \sup(F(K))$.

6 Fuzzy Ontology Generation

This step constructs fuzzy ontology from a fuzzy context using the concept hierarchy created by fuzzy conceptual clustering. This is done based on the characteristic that both FCA and ontology support formal definitions of concepts.

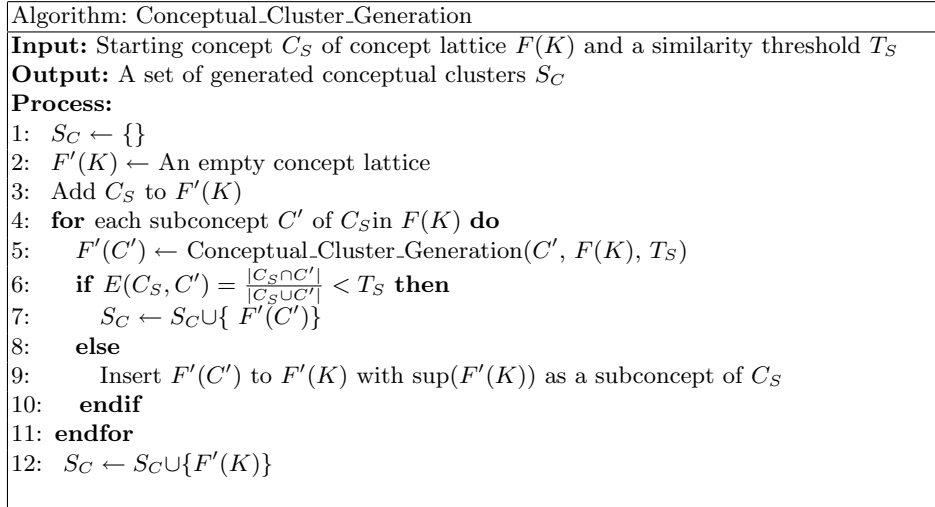


Fig. 6. The fuzzy conceptual clustering algorithm.

However, a concept defined in FCA has both extensional and intensional information [3], whereas a concept in an ontology only emphasizes on its intensional aspect. To construct the fuzzy ontology, we need to convert both intensional and extensional information of FCA concepts into the corresponding classes and relations of the ontology.

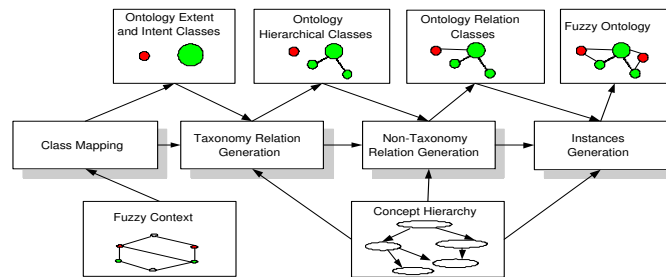


Fig. 7. Fuzzy ontology generation process

Figure 7 shows the ontology generation process, which consists of the following steps.

- *Class Mapping* – It maps the extent and intent of the fuzzy context into the extent and intent classes of the ontology. Human interpretation is required to label the extent class name. For example, the extent class mapped from

the extent of the fuzzy context given in Table 1(b) can be labeled manually as *Document*. We can use appropriate names to represent keyword attributes and use them to label the intent class names as well. For example, the class *Research Area* can be used to label the initial intent class.

- *Taxonomy Relation Generation* – It expands the intent class of the ontology as a hierarchy of classes using the concept hierarchy. The process can be considered as an isomorphic mapping from the concept hierarchy into taxonomy classes of the ontology. For example, the class *Research Area* can be expanded into a hierarchy of classes, in which each class represents a research area, corresponding to the concept hierarchy given in Figure 3(b).
- *Non-taxonomy Relation Generation* – It generates the relation between the extent class and intent classes. This task is quite straightforward. However, we still need to label the non-taxonomy relation. For example, the relation between the class *Document* and class *Research Area* can be labeled as *belong-to* relation.
- *Instances Generation* – It generates instances of the extent class. Each instance corresponds to an object in the initial fuzzy context. Based on the information available on the fuzzy concept hierarchy, instances' attributes are automatically furnished with appropriate values. For examples, each instance of the class *Document* (which corresponds to an actual document) will be associated with the appropriate research areas.

7 Scholarly Ontology for Scholarly Semantic Web

We have applied the Fuzzy Ontology Generation framework to generate the Scholarly Ontology for Scholarly Semantic Web from a citation database. The citation database is created from a set of 1400 scientific documents on the research area “Information Retrieval” published in 1987-1997 downloaded from Institute for Scientific Information’s (ISI) [13]. The construction of fuzzy formal context is done as follows. For each document, we have extracted the 10 most frequent citation keywords. We then construct a fuzzy formal context $K_f = \{G, M, I\}$, with G as the set of documents and M as the set of keywords. The membership value of a document D on a citation keyword C_K in K_f is computed as

$$\mu(D, C_K) = \frac{n_1}{n_2}$$

where n_1 is the number of documents that cited D and contained C_K and n_2 is the number of documents that cited D . This formula is based on the premise that the more frequent a keyword occurs in the citing paper, the more important the keyword is in the cited paper.

The FOGA framework is then applied to the fuzzy formal context. We obtain a set of intent and extent classes of the ontology. We label the extent class of the ontology as *Document*. The generated intent classes represent *research areas* of documents. The ontology taxonomy and non-taxonomy relations between

ontology classes are automatically generated. However, we need to label the non-taxonomy relation between the extent and intent classes. This relation is labeled as *belong-to* relation, which implies that “*a document belongs to research areas*”.

As such, the scholarly ontology generated for the Scholarly Semantic Web contains scholarly information as a hierarchy of research areas as well as research areas for each document. Figure 8 depicts a part of the generated concept hierarchy of research areas. We use the keyword that has the highest membership value to label the research area. Nevertheless, users can browse more detail information of each research area.

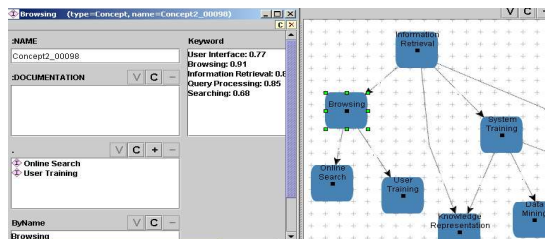


Fig. 8. A part of the concept hierarchy of research areas

8 Performance Evaluation

The scholarly ontology is generated based on the concept hierarchy constructed automatically by FOGA. Therefore, the quality of the ontology generated depends directly on the concept hierarchy generated. To evaluate the concept hierarchy, we use the *relaxation error* (RE) [14] to measure the goodness of the concepts generated. In addition, we also measure the *Average Uninterpolated Precision* (AUP) [15] to evaluate the retrieval performance from the concept hierarchy.

8.1 Evaluation Using Relaxation Error

To evaluate the goodness of the clusters generated, we measure the relaxation error, which implies dissimilarities of items in a cluster based on attributes' values. The relaxation error RE of a cluster C is defined as

$$RE(C) = \sum_{a \in A} \sum_{i=1}^n \sum_{j=1}^n P(x_i) P(x_j) d^a(x_i, x_j)$$

where A is the set of attributes of items in C , $P(x_i)$ is the probability of item x_i occurring in C and $d^a(x_i, x_j)$ is the distance of x_i and x_j on attribute a . In

our application, $d^a(x_i, x_j) = |m(i, a) - m(j, a)|$ where $m(i, a)$ and $m(j, a)$ are the membership values of objects x_i and x_j on attribute a respectively. The cluster goodness G of cluster C is defined as

$$G(C) = 1 - RE(C)$$

Since COBWEB is considered as one of the most popular techniques for conceptual clustering, we also apply COBWEB to the citation database to compare the performance. To use COBWEB, the membership values of keywords are replaced by appropriate nominal values. If the membership value is greater than 0.5, it is set as “Yes”, otherwise it is set as “No”.

Figure 9 shows the performance evaluation results on cluster goodness using FOGA and COBWEB while the number of extracted keywords is varied from 2 to 10. As shown in Figure 9, FOGA has achieved better cluster goodness than COBWEB. This has demonstrated the advantage of using fuzzy membership values for representing object attributes. In addition, the experimental results have also shown that good cluster goodness is obtained when the number of extracted keywords is small. It is expected because smaller number of keywords used will cause smaller differences in objects in terms of keywords’ membership values. Therefore, the relaxation error will be smaller. However, as we will see later in Section 8.2, smaller number of extracted keywords will cause poor retrieval performance.

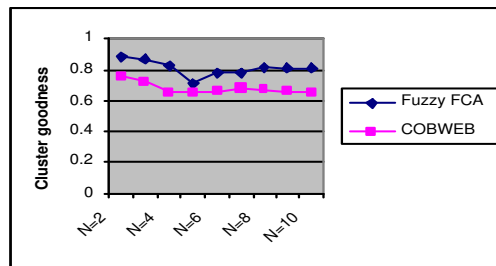


Fig. 9. Performance evaluation results on cluster goodness

8.2 Evaluation Using Average Uninterpolated Precision

The Average Uninterpolated Precision (AUP) is defined as the sum of the precision value at each point (or node) in a hierarchical structure where a relevant item appears, divided by the total number of relevant items. For evaluating AUP, we have manually classified the downloaded documents into classes based on their research themes. For each class, we extract 5 most frequent keywords from the documents in the class. Then, we use these keywords as inputs to

form retrieval queries and evaluate the retrieval performance using AUP. This is carried out as follows. For each document, we will generate a set of *document keywords*. There are two ways to generate document keywords. The first way is to use the set of keywords, known as *attribute keywords*, from each conceptual cluster as the document keywords. The second way is to use the keywords from each document as the document keywords. Then, we vectorize the document keywords and the input query, and calculate the vectors' distance for measuring the retrieval performance.

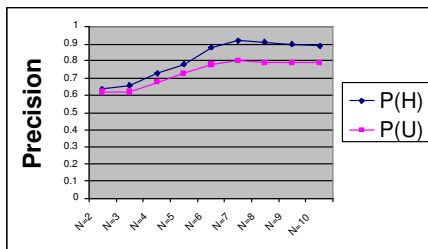


Fig. 10. Performance evaluation on precision.

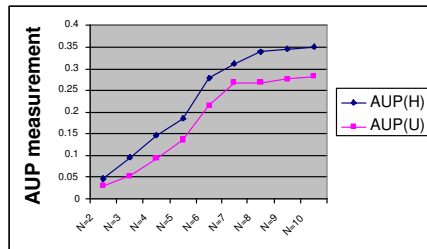


Fig. 11. Performance evaluation on AUP.

We refer the precision and AUP measured using the first way (i.e. using attribute keywords) to as *Hierarchical Precision* ($P(H)$) and *Hierarchical Average Uninterpolated Precision* ($AUP(H)$), as each concept inherits attribute keywords from its superconcepts. Whereas the precision and AUP measured using the second way (i.e. using keywords from documents) is referred to as *Unconnected Precision* ($P(U)$) and *Unconnected Average Uninterpolated Precision* ($AUP(U)$).

Figures 10 and 11 give the performance results for $P(H)$ and $P(U)$ and $AUP(H)$ and $AUP(U)$ using different numbers of extracted keywords N . From Figure 11, we found that when N gets larger, the performance on $P(H)$, $P(U)$, $AUP(H)$ and $AUP(U)$ gets better. When N is larger than 5, the values of $P(H)$ and $P(U)$ are considered as good performance (around 0.9 and 0.8 respectively). It has shown that the number of keywords extracted for conceptual clustering has affected the retrieval performance. In addition, the performance results on $P(H)$ and $AUP(H)$ are generally better than that of $P(U)$ and $AUP(U)$ respectively. It implies that the attribute keywords generated for conceptual clusters are more appropriate concepts for representing the concept hierarchical structure.

9 Conclusions

In this paper, we have proposed the FOGA framework for fuzzy ontology generation on uncertainty information. FOGA consists of the following steps: Fuzzy Formal Concept Analysis, Fuzzy Conceptual Clustering and Fuzzy Ontology

Generation. In addition, the FOGA framework has been applied to generate scholarly ontology for the Scholarly Semantic Web from citation databases. The generated scholarly ontology represents knowledge on documents and its research areas. The performance evaluation of the proposed FOGA framework has also been given based on the generation of the scholarly ontology.

References

1. N. Guarino and P. Giaretta, *Ontologies and Knowledge Bases: Towards a Terminological Clarification. Toward Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, IOS Press, Amsterdam, 1995.
2. D.H.Widyantoro and J.Yen, "A Fuzzy Ontology-based Abstract Search Engine and Its User Studies", In *Proceedings of the 10th IEEE International Conference on Fuzzy Systems*, 2001, Melbourne, Australia, 2001, pp. 1291-1294.
3. B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*, Springer, Berlin – Heidelberg, 1999.
4. S. Kampa, T. Miles-Board and L. Carr, "Hypertext in the Semantic Web", In *Proceedings ACM Conference on Hypertext and Hypermedia*, Aarhus, Denmark, 2001, pp. 237-238.
5. A. Maedche and S. Staab, *Ontology Learning for the Semantic Web*. IEEE Intelligent Systems, Special Issue on the Semantic Web, Vol. 6, No.2, 2001.
6. A. Faatz and R. Steinmetz, "Ontology enrichment with texts from the WWW", in *Proceedings of Semantic Web Mining second Workshop at ECML/PKDD-2002*, Finland, 2002.
7. G. Bisson and C. Nedellec, "Designing Clustering Methods for Ontology Building: The Mo'K Workbench", In S. Staab, A. Maedche, C. Nedellec, P. WiemerHasting, editors, In *Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence, ECAI'00*, Germany, 2000.
8. P. Clerkin, P. Cunningham and C. Hayes, "Ontology Discovery for the Semantic Web Using Hierarchical Clustering", in *Proceedings of Workshop at ECML/PKDD-2001*, Germany, 2001.
9. G. Stumme, R. Taouil, Y. Bastide, N. Pasquier and L. Lakhan, "Computing iceberg concept lattice with Titanic", *Journal on Knowledge and Data Engineering*, Vol. 42, No. 2, 2002, pp. 189-222.
10. F. Vogt and R. Wille, "TOSCANA: a Graphical Tool for Analyzing and Exploring Data", In R.Tamassia and I. G. Tollis, editors, *GraphDrawing' 94*, Heidelberg, 1995, pp.226-233.
11. S. Pollandt, *Fuzzy-Begriffe: Formale Begriffsanalyse unscharfer Daten*, Springer Verlag, Berlin – Heidelberg, 1996.
12. L.A Zadeh, "Fuzzy Sets", *Journal of Information and Control*, Vol. 8, 1965, pp. 338-353.
13. ISI, *Institute for Scientific Information*, Available at: <<http://www.isinet.com>>, 2000.
14. W. Chu and K. Chiang, "Abstraction of High Level Concepts from Numerical Values in Databases", In *Proceedings of AAAI Workshop on Knowledge Discovery in Databases*, 1994, pp. 133-144.
15. N.Nanas, V.Uren and A. de Roeck, "Building and Applying a Concept Hierarchy Representation of a User Profile", In *Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, 2003.