# Methodology of Ontology Extraction for Islamic Knowledge Text

Saidah Saad,
Department of Information System,
Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia, Skudai, Johor.
603-89216717
saidah@ftsm.ukm.my,

Naomi Salim
Department of Information System,
Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia, Skudai, Johor.
607-5532211
naomie@fsksm.utm.my

## ABSTRAK

Ontology plays an essential role in the formalization of information for effective human-computer interactions. However, engineering of domain ontologies turns out to be very labor intensive and time consuming. Recently, some machine learning methods have been proposed for automatic discovery of domain ontologies. Nevertheless, the accuracy and computational efficiency of the existing methods need to be improved to support large scale ontology construction for real-world applications. This paper illustrates an ontology extraction based on fuzzy-swarm algorithm for Islamic knowledge (IK) text. By combining lexico-syntactic and statistical learning methods, the accuracy and the computational efficiency of the ontology discovery process is improved. Empirical studies have confirmed that the proposed method can discover high quality fuzzy domain ontology which leads to significant improvement in information retrieval performance.

## Keyword:

Keyphrase, FFCA, Clustering, Hybrid PSO, Ontological Component.

## 1. INTRODUCTION

Internet technology has made IT users aware of both new opportunities as well as actual needs for large scale interoperation of distributed, heterogeneous, and autonomous information systems. Additionally the vastness of the amount of information already on-line, or to be interfaced with the WWW, makes it unfeasible to depend merely on human users to correctly and comprehensively identify access, filter and process the information relevant for the purpose of applications over a given domain. Be they called software agents, web services, or otherwise, this is increasingly becoming the task of computer programs equipped with domain knowledge. Presently however there is an absence of usable formal, standardized and shared domain knowledge of what the information stored inside these systems and exchanged through their interfaces actually means. Nevertheless this is a prerequisite for agents and services (or even for human users) wishing to access the information but who, obviously, were never involved when these systems were created. The pervasive and explosive proliferation of computerized information systems (databases, intranets, communication systems, or other) quite simply makes this into the key problem of the application layer of the current internet and its semantic web successor. The equally obvious key to the solution of this problem therefore lies in a better understanding, control and management of the semantics of information in a general sense. The semantic principles and technology underlying such solutions are emerging in the form of ontologies.

The manual construction of this ontologies will require enormous human efforts, thus ontology acquisition becomes a bottleneck of the Semantic Web. Now, ontologies are a popular research topic in various communities such as knowledge engineering, natural language processing, cooperative information systems, intelligent information integration, and knowledge management. They provide a shared and common understanding of a domain that can be communicated between people and heterogeneous and distributed application systems. They have been developed in Artificial Intelligence to facilitate knowledge sharing and reuse[8].

Modern research focus lies in Web-based ontology representation languages based on XML and RDF standards and further application of ontologies on the Web (Decker et al., 2000). Ontology learning (OL) is an emerging field aimed at assisting a knowledge engineer in ontology construction and semantic page annotation with the help of machine learning (ML) techniques.

## 2. KEYPHRASE EXTRACTION

Keyphrases for a document are useful for many applications. For text retrieval keyphrases can help narrow search results or rank retrieved documents. They can be used to cluster semantically related documents for the purposes of categorization.

In our research, we propose to exploit a keyword and keypharse extraction in order to identify relevant terms in the document. It is because, most of the concept and attributes of the concept occur in Islamic Knowledge, being express in phrase.

According to Huang, C. et. al [5], they are two-phase filtering algorithm for extracting keyphrase more effective. First, phrases must fulfill three rules, (i). A phrase can't start or end up with stopwords, (ii) in a phrase, only a word sequence of less than four midwords (propositions, nouns, numbers, and some conjunctions in stopword list) can exist between two non-

stopwords. Phrases as "members of the family" are included, (iii) frequency of a phrase (PF) is above a minimum value. Second, we select phrases with relatively higher PFs.

They are first divided into societies by the number of nonstopwords. Then phrases with the same word in a society are further assigned to the same group. A phrase is then in n groups if it has n distinct non-stopwords. Every group has only one or none winner. A winning keyphrase candidate (named a giant phrase) should have a top PPF (Percent of PF) and a PTF (Percent of TF) above a threshold in all n groups it belongs to. Finally, only winners can remain.

$$PTF(phrase_i, group_k) = \frac{PF_i}{TF_k}$$

$$PPF(phrase_i, group_k) = \frac{PF_i}{\sum_{phrase_j \in group_k} PF_j}$$

## 3. ONTOLOGICAL COMPONENT

Ontological components that had been proposed by Omelayenko[10] with additional component (with gray color box) is design in order of querying process is presented in Figure 1. First, the user formulates the query in natural language. Then the query is transformed into a formal query with the help of the natural language ontology and the domain ontology. The documents are (possibly incomplete) instances of some domain ontologies, and they will contain pieces of data semantically marked up according to the underlying domain ontology. The query processor has to find the mapping between the concepts of the initial query, the domain model used to expand the query, and the ontology instances on the document. This mapping will be non-trivial and will require inference over domain ontologies. The gold standard will use to evaluate methods for ontology population on the instance level where the knowledge engineer will annotate instances in the text with the appropriate concept from the ontology.

### 3.1 Natural Language Ontology (NLO) Creation

NLO contain lexical relations between the language concepts (words and their senses). They usually represent as background knowledge of the system. They try to capture all possible concepts, but they do not provide detail description of each of the concepts. General language knowledge contained in general purpose Natural Language Ontology (NLO) like WordNet [7] will be used to link text to specific terms and concepts. The system that had been generated exploits the text from the documents to enrich the concepts in the wordNet. In Islamic Knowledge document, there are terms that have different meaning from what in wordNet and they have their own unique term/phrase. This NLO will be used to enrich all possible language concepts that already had in WordNet plus with unique term/phrase in Islamic Knowledge. The keyphrase technique will be explored in order to generate meaningful concept that relevant to Islamic Knowledge.

In learning NLO, lot of conceptual clustering methods had already been used for ontology construction by several researchers such as Bisson et. al.[3]; Agirre et. al. [1]; Faure & Poibeau, [6], Quan et. al.[13] and Cui & Potok. [4].

For Islamic Knowledge NLO, we will develop using Fuzzy Formal Concept Analysis (FFCA) [12], [14]. Ontology constructed by formal concept analysis is quite complicated in terms of the number of concepts generated and cannot deal with the vague and uncertain information in practice. We also explores the role of swarm intelligent (SI) in clustering different kinds of datasets and a new SI technique for partitioning any dataset into an optimal number of groups through one run of optimization will be proposed.

The whole process from FFCA to domain ontology is shown in figure 2.

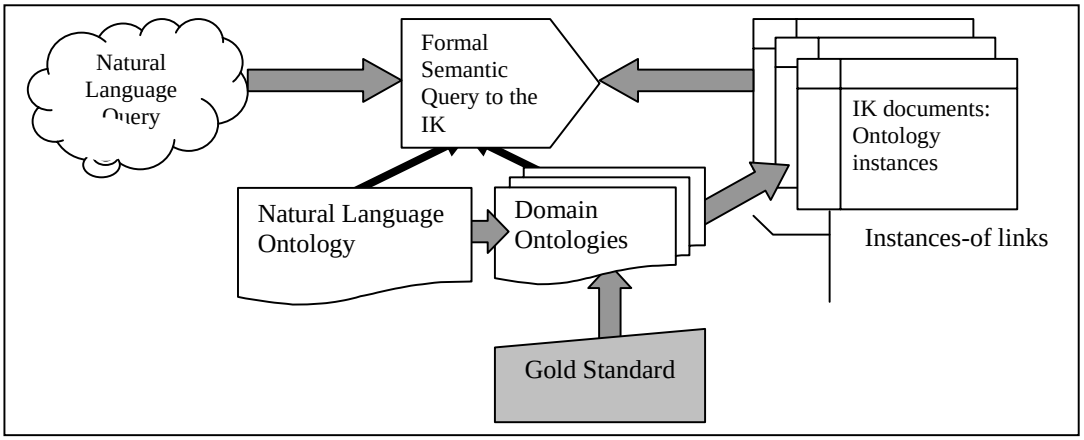


Figure 1. Semantic querying of the Web

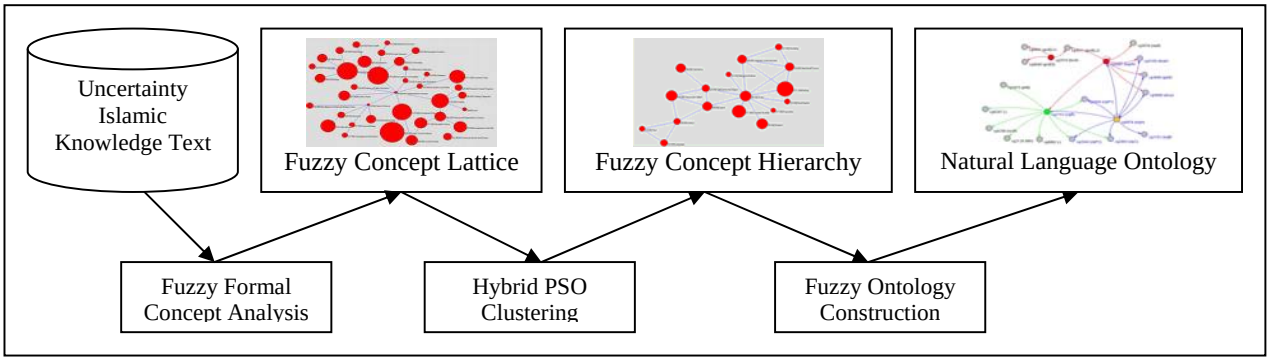


Figure 2. The process of ontology learning by SI clustering

The figure 2 above shows the proposed architecture which consist of the following components:

## *i. Fuzzy Formal Concept Analysis (FFCA)*

FFCA incorporates fuzzy logic into formal concept analysis to represent vague information of Islamic Knowledge. The definitions of FFCA are (Quan, et.al. 2006)

Definition 1 : A fuzzy formal context is a triple K = (G, M, I = φ(G £ M)) where G is a set of objects, M is a set of attributes, and I is a fuzzy set on domain G x M. Each relation (g,m) ∈ I has a membership value μ(g;m) in [0,1].

A confident threshold T can be set to eliminate relations that have low membership values.

Definition 2: Given a fuzzy formal context K=(G, M, I) and a confidence threshold T, we define A*= {m ∈ M|∀g ∈ A: μ (g, m) ≥ T} for A ⊆ G and B*= {g ∈ G|∀m ∈ B: μ (g,m) ≥ T} for B ⊆ M. A fuzzy formal concept (or fuzzy concept) of a fuzzy formal context (G, M, I) with a confidence threshold T is a pair (Af =φ (A), B) where A ⊆ G, B ⊆ M, A* = B and B* = A: Each object g ∈ φ (A) has a membership μg defined as

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

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

Definition 3: Let (A1, B1) and (A2, B2) be two fuzzy concepts of a fuzzy formal context (G, M, I). (φ(A1), B1) is the subconcept of (φ(A2), B2), denoted as (φ(A1), B1)≤ (φ(A2), B2), if and only if φ(A1) ⊆ φ(A2)(⇔ B2 ⊆ B1). Equivalently, (A2, B2) is the superconcept of (A1, B1).

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 ≤ with the confidence threshold T.

## *ii. Hybrid PSO Clustering*

Hybrid Particle Swarm Optimizers is combining the idea of the particle swarm with K-means algorithm because, although the PSO algorithm generates much better clustering result than K-menas algorithms does, in term of execution times, K-means algorithms is more efficient for large dataset and the previous research showed that the hybrid PSO algorithm can generate more compact clustering result.[4][11] .

The Hybrid PSO algorithm can be summarized as [4][11] :

Start the PSO clustering process until the maximum number of iterations is exceeded.

(1) At the initial stage, each particle randomly chooses k different document vectors (fuzzy value) from the document collection as the initial cluster centroid vectors.

(2) For each particle:
   (a) Assign each document vector in the document set to the closest centroid vector.

   (b) Calculate the fitness value based on equation below

$$f = \frac{\sum_{i=1}^{N_c} \{ \frac{\sum_{j=1}^{p_i} d(o_i, m_{ij})}{p_i} \}}{N_c}$$

(c) Using the velocity and particle position to update equations (a) and (b) below and to generate the next solutions.

$$v_{i(}d+1) = w \times v_{i(}d) + c_1 \times \left(p_{i(}d) - x_{i(}d)\right) + c_2 \times \left(p_{g(}d) - x_{i(}d)\right) \rightarrow \text{(a)}$$

$$x_{i(}d+1) = x_{i(}d) + v_{i(}d+1) \rightarrow \text{(b)}$$

(3) Repeat step (2) until one of the following termination conditions is satisfied.
   (a) The maximum number of iterations is exceeded or
   (b) The average change in centroid vectors is less than a predefined value.

Start K-Means process until maximum number of iterations is reached.

(1) Inherit clustering result from PSO as the initial centroid vectors of K-means module.
(2) Assigning each document vector to the closest cluster centroids.
(3) Recalculating the cluster centroid vector cj using equation below.

$$c_j = \frac{1}{n_j} \sum_{\forall d_j \in S_j} d_j$$

where dj denotes the document vectors that belong to cluster Sj; cj stands for the centroid vector; nj is the number of document vectors belong to cluster Sj.

(4) Repeating step 2 and 3 until the convergence is achieved.

### *iii. Fuzzy NLO Construction*

This step constructs Fuzzy NLO from fuzzy context using the concept hierarchy created by hybrid PSO clustering. This is done based on the characteristic that both FCA and ontology support formal definitions of concepts. However, a concept defined in FCA has both extensional and intentional information, whereas a concept in an ontology only emphasizes on its intentional aspect. To construct the fuzzy ontology, we need to convert both intentional and extensional information of FCA concepts into the corresponding classes and relations of the ontology.

The following step is the ontology construction process proposed by Quan et. al.[13]

- Class Mapping
  It maps the extent and intent of the fuzzy context into the extent and intent classes of the ontology. For example, the extent class can mapped from the extent of the fuzzy context and use appropriate names to represent keyword attributes and use them to label the intent class names as well.

- 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.

- 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 in stance 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 in stance of the class Document (which corresponds to an actual document) will be associated with the appropriate research areas.

## 3.2 Islamic Knowledge Gold Standard Creation

This task will include collection of initial corpus containing terminology considered by the domain experts as highly relevant for the ontology. According to Al Kabi et al,[2], the general subjects that they found in Quran and Hadith that are relevant to Muslim scholars' classifications are Islamic basic (Islam Pillars), Faith, General and political relations, Science and art, Holy quran, Organizing financial relations, Human and social relations, Al-Jehad, Religions, Judicial relations, Working, Stories and history, Human and ethical relations, Trade, agriculture and industry and Call for Allah (Dakwah).

The gold standard developed will be based on these general subjects as a guide. Since it is anticipated that Islamic domain ontology are complex in structure, the gold standard will be construct manually and focus be first made on solat-related ontology.

## 3.3 Islamic Knowledge Domain Ontology Creation

For this task, the Islamic Knowledge Text, NLO and gold standard will be use as an input and generate ready-to-use

Islamic ontology as output with the possible help of knowledge engineer. The process of ontology FFCA merging and pruning which described the three steps of the technique: the linguistic analysis of the texts which returns two formal contexts; the merging of the two contexts and the computation of the pruned concept lattice[15] will be applied here.

### 3.4 Islamic Knowledge Ontology Instances Creation

This is commonly referred to as Ontology Population. The task of populating an ontology is very related to the named entity recognition in extracting instances from Islamic Knowledge text. However the technique that mention before (FFCA) also will be used in order to extract instances from the document text.

## 4 CONCLUSION

In this paper, we proposed the design of FFCA with hybrid PSO clustering for methodology of Ontology Extraction for Islamic Knowledge Text. First we describe the main key in developing this ontology which is, how to extract concept and attribute in order to develop FFCA using keyword and keyphrase. Second, the ontological component which are successfully applied by previous research such as Natural Language Ontology (NLO); Islamic Knowledge Gold Standard; Domain Ontology; Ontology Instances. The performance evaluation of the proposed design methodology will be generating in several stages in future in order to confirm the design.

## 5 REFERENCES


[1]. Agirre, Ansa, Hovy, Martinez: 2000. Enriching very large ontologies using the WWW. In: S. Staab, A. Maedche, C. Nedellec, P. Wiemer-Hastings (eds.), Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI'00, Berlin, Germany, August 20-25,

[2]. AL-Kabi M.N. Kanaan G. And Al-Shalabi R. 2005. Statistical Classifier of the Holy Quran Verses (Fatiha and Yaseen Chapters). Journal of Applied Sciences 5(3): 580-583,

[3]. Bisson, Nedellec, Canamero. 2000.: Designing Clustering Methods for Ontology Building - The Mo'K Workbench. In: S. Staab, A. Maedche, C. Nedellec, P. Wiemer Hastings (eds.), Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI'00, Berlin, Germany, August 20-25,

[4]. Cui X and Potok TE, 2005. Document Clustering Analysis Based on Hybrid PSO+Kmeans Algorithm, Journal of Computer Sciences (Special Issue), ISSN 1549-3636, pp. 27-33.

[5]. Chong Huang, Yonghong Tian, Zhi Zhou, Charles X. Ling,Tiejun Huang. 2006. Keyphrase extraction using Semantic Networks Structure Analysis. In Procceding of the sixth IEEE International Conference on Data Mining (ICDM 2006), Hong Kong, ,pp. 275-284, IEEE press.

[6]. Faure, Poibeau. 2000. First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. In: S. Staab, A. Maedche, C. Nedellec, P. Wiemer-Hastings (eds.), Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI'00, Berlin, Germany, August 20-25,

[7]. Fellbaum. 1998. WordNet: An Electronic Lexical Database. The MIT Press,

[8]. Fensel. 2000 .Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce. Springer-Verlag, Berlin,

[9]. Decker, Fensel, Van Harmelen, Horrocks, Melnik, Klein, Broekstra: 2000. Knowledge Representation on the Web. In: Proceedings of the 2000 International Workshop on Description Logics (DL2000), Aachen, Germany, August,

[10]. Omelayenko 2001.. Learning of ontologies for the Web: the analysis of existent approaches. In Proceedings of the International Workshop on Web Dynamics,

[11]. Omran, Engelbrecht and Salman. 2005. Dynamic Clustering using Particle Swarm Optimization with Application in Unsupervised Image Classification. Fifth World Enformatika Conference (ICCI 2005), Prague, Czech Republic,

[12]. Quan, Hui, Fong, Cao, 2006.: Automatic fuzzy ontology generation for Semantic Web. IEEE Transactions on Knowledge and Data Engineering 18(6), 842–856

[13]. Quan, Hui, and Cao, 2004. “FOGA: A Fuzzy Ontology Generation Framework for Scholarly Semantic Web”, Proceedings of the 2004 Knowledge Discovery and Ontologies Workshop (KDO’04), Pisa, Italy,

[14]. Zhou, Liu, Zhou, 2007. "Ontology Learning by Clustering Based on Fuzzy Formal Concept Analysis," compsac, pp. 204-210, 31st Annual International Computer Software and Applications Conference - Vol. 1- (COMPSAC 2007),

[15]. Stumme and Madche. 2001. FCA-Merge: Bottom-up merging of ontologies. In 7th Intl. Conf. on Artificial Intelligence (IJCAI '01), pages 225--230, Seattle, WA,