Design of an Automatic Ontology Construction Mechanism using Semantic Analysis of the Documents

Prashant Dixit
Department of Information Technology
Manav Rachna International University
Faridabad, Haryana, India
prashantdixit2k6@gmail.com

Shilpa Sethi, A. K. Sharma, Ashutosh Dixit
Department of Computer Engineering
YMCA University of Science & technology
Faridabad, Haryana, India
munjal.shilpa@gmail.com, ashokkale2@rediffmail.com,
dixit_ashutosh@rediffmail.com

Abstract- **Ontologies play a major role in supporting information exchange processes in various areas. At present, ontologies are applied to the World Wide Web for creation of semantic web. The main application area of ontology technology is Knowledge Management. In the present scenario, it is difficult to acquire knowledge and then to maintain knowledge in a given domain. Manual ontology population is labor intensive and time consuming. Hence there is need to devise a method to provide fully automatic feeding of Web-based knowledge to the ontology. Moreover, for constructing ontology automatically, there is a need to discover a way to find, structure, and display the relationships between attributes and objects of a sentence. In this paper, a technique of Automatic Semantic Domain-ontology Populator (ASDP) towards construction of given domain modeled by the database is being proposed. ASDP is a way to find, structure, and display relationships between concepts, which consist of attributes and objects. This method helps in understanding a given domain and in building a domain model for it.**

Index Terms - Ontology, Domain-ontology, Information Retrieval.

I. INTRODUCTION

Ontologies play a major role in supporting information exchange processes in various areas [1, 5, 6]. In the past, ontologies were developed in Artificial Intelligence to facilitate knowledge sharing and reuse. Since the beginning of the nineties, ontologies have become a popular research topic investigated by several Artificial Intelligence research communities, including Knowledge Engineering, Natural Language Processing and Knowledge Representation. Recently, the idea of ontology is also becoming common in fields such as intelligent information integration, cooperative information systems, information retrieval, electronic commerce, and knowledge management. The main reason behind the ontologies becoming popular is that they assure a shared and common understanding of some domain that can be communicated between people and application systems.

II. LITERATURE SURVEY

In the following section some of the existing ontology construction methods are discussed.

A. TERMINAE

The purpose of TERMINAE [1] is to build automatic ontology from text as well as a new ontology manually. It is a computer aided Knowledge-Engineering tool written in java. TERMINAE is composed of following two tools.

- Linguistic Engineering Tool
- Knowledge Engineering Tool

Linguistic Engineering Tool: This module allows the extraction of terminological forms (keywords) from the given corpus (Text file). Terminological forms define each meaning of a term called a notion using some linguistic relation (Parts-of-Speech) between notions such as synonyms.

Knowledge Engineering Tool: This module involves knowledge base (Ontology) management with an editor and browser for the ontology. The tool helps to represent a notion (topic or keyword) as a concept. This tool can directly be used to create the ontology from scratch.

B. Ontology Development using SALT [2]

It is the common idea of two different projects: The standardization of lexical and terminological resources (SALT) and the use of conceptual ontologies for information extraction and data integration (TIDIE). This approach assumes the availability of following 3 types of knowledge sources

- More general and well defined ontology for the domain.
- A dictionary or any external source to discover lexical and structural relationships like WordNet.
- Consistent set of training text documents.

And to extract Ontology knowledge source must

- Be of a general nature.
- Contain meaningful relations.
- Already exist in Machine readable form.
- Have a straight forward conversion into XML.

The architectural view of SALT is given in Fig. 1

978-0-7695-4850-0/12 $26.00 © 2012 IEEE
DOI 10.1109/CICN.2012.89

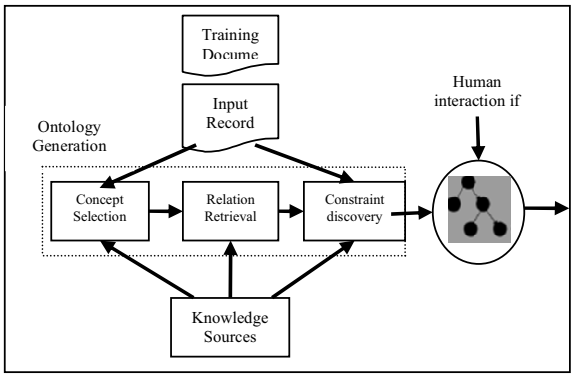


Figure 1: Ontology generation process of SALT

The different modules of SALT are described below:

Concept selection: Select the user required concepts from the domain. This is done by string matching between textual content and ontological data.

Relationship retrieval: First find out the conceptual relationships from the knowledge sources. Now construct a directed graph whose nodes are concepts. And the relationships between these concepts can be represented by paths among the concepts. To find the relationships more accurately use Dijkstra's algorithm, to find out the shortest path (more appropriate) relations among the concepts.

Constraint Discovery: constraints such as a person can have only one Date of Birth, two parents and several phone numbers follows adopted conventions.

Refining results: The output ontology may not be the final ontology which user can directly use. An expert will revise and refine the ontology.architectural view

C. Ontology Construction for Information Selection [4]

The procedure explained for ontology construction for information retrieval explained by Latifur Khan and Feng Luo is given below

- Terms are extracted from documents with text mining techniques.
- Documents are grouped hierarchically according to their similarities using a modified version of SOTA algorithm.
- Assign concepts to the tree nodes starting from leaf nodes with a method based on the Rocchio algorithm.
- Concept assignment is based on WordNet hyponyms.
- Bottom up approach for ontology generation.

A comparison between the existing ontology construction techniques is given in Table 1.

TABLE I.

COMPARISON OF ONTOLOGY CONSTRUCTION METHODS

	Extraction	Analysis	Generation	Validation
TERMINAE	NLP tools are used, Human intervention is optional	Concept Relationship analysis (Semi-automated)	No standard Ontology representation	Purely by human
SALT	NLP Techniques fully automated	Similarity analysis of concepts	No standard Ontology representation	Limited human intervention
Ontology Construction for Information selection	Human intervention is optional	Not provided	Human intervention optional	Not provided

III PROBLEM IDENTIFICATION

A critical look at the available literature indicates the following problems in the present knowledge management system:

- It is difficult to acquire knowledge and then to maintain knowledge in a given domain.
- Manual ontology population is labour intensive and time consuming. In the current scenario no direct input available to the ontology. Hence there is need to devise a method to provide fully automatic feeding of Web-based knowledge to the ontology.
- For constructing ontology automatically, there is a need to discover a way to find, structure, and display the relationships between attributes and objects of a sentence.

IV PROPOSED WORK

In this work, a technique of Automatic Semantic Domain-ontology Populator (ASDP) towards construction of given domain modelled by the database is being proposed. ASDP is a way to find, structure, and display relationships between concepts, which consist of attributes and objects. This method helps in understanding a given domain and in building a domain model for it. Automatic Semantic Domain-ontology Populator (ASDP) is a method for deriving conceptual structures out of data. These structures can be represented as XML file [7].

ASDP based on the philosophical understanding that a concept is constituted by two parts: its extension which consists of all thematic objects belonging to the concept, i.e. ability of the system to span over all thematic objects belonging to the concepts and its intension which comprises all attributes shared by those thematic objects i.e. to include or exclude the objects having common attribute or characteristics. This understanding allows to derive all concepts from a given context.

It has been identified that the five phases are required for the automatically extracting the semantic information from the text documents and populate them in ontology.

1. **Lexical Processing**: ASDP starts with the lexical processing of a plain text. The sub phases if lexical analysis phase are sentence separation, tokenize, part of speech labelling, and morphological analysis.

2. **Syntactic Processing**: In this phase English grammar rules are applied to the input received from lexical processing phase.
3. **Anaphora Resolution** :The process of anaphora Resolution is the problem of resolving what a pronoun or a noun phrase refers to that is previously defined in the document
4. **Semantic Processing:** This phase helps in identifying the main parts of a sentence i.e. object, subject, actions, attributes.
5. **Ontology Population:** In this phase, the information is inserted into the Knowledge Base following the ontology domain representation.

The block diagram of the proposed system is given in Fig. 2.

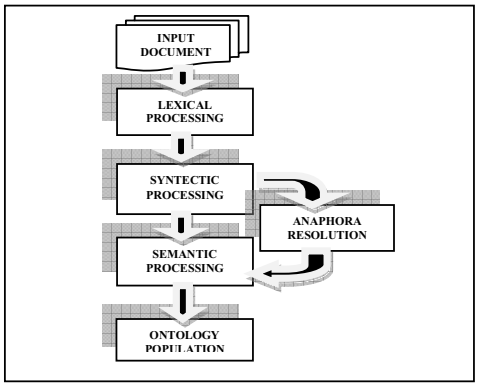


Figure 2. Block diagram of Automatic Semantic Domain-ontology Populator (ASDP)

A detailed discussion on the each phase is given in the following section.

A. Lexical Processing

The Natural Language parsing starts with the lexical processing of a plain text. The lexical processing phase comprises following four sub-phases as shown in Fig. 3.

1. Sentence Separation
2. Tokenize
3. Parts-of-Speech labelling
4. Morphological Analysis

The description of each sub field is given below:

1) Sentence separation

The statements in text document are further processed to identify the margins of a sentence and split as individual sentences using the stop points. The output of this module is an array of individual sentences to be considered for further processing. Each of the individual sentences is given as an input to the next module.

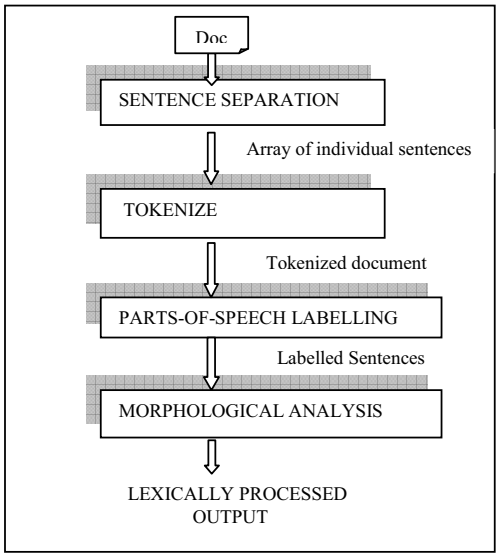


Figure 3. Lexical Processing

2) Tokenization

In tokenization part, the sentences are divided in to its component words called Tokens. For example "Customer has purchased a red ball." is tokenized as [Customer] [has] [purchased] [a] [red] [ball] [.]. It takes the input from sentence separation module and provides output to part of speech labelling module.

3) Parts-Of-Speech(POS) Labelling

In this phase each sentence is further passed to Stanford parts-of- speech (POS) tagger v3.0 to identify the basic POS tags. The Stanford POS tagger v3.0 can identify 44 types of POS tags. For example, the tokenised sentence [Customer] [has] [purchased] [a] [red] [ball] [.] is labelled as [Customer/NN] [has/HV] [purchased/VB] [a/DT] [red/JJ] [ball/NN] [.]

4) Morphological Analysis

Migrant Words are potentially complex units, composed of even more basic units, called morphemes. A morpheme is the smallest part of a word that has grammatical function or meaning. In this work, the morphemes are designate in braces—{ }. For example, *sawed*, *sawn*, *sawing*, and *saws* can all be analyzed into the morphemes {saw} + {-ed}, {-n}, {-ing},and {-s}, respectively.

B. Syntactic Processing

In this phase English grammar rules are applied to the input received from lexical processing phase. The text is syntactically analyzed and a parse tree is generated for further semantic processing. A broad level view of syntactic processing phase is given in Fig. 4.

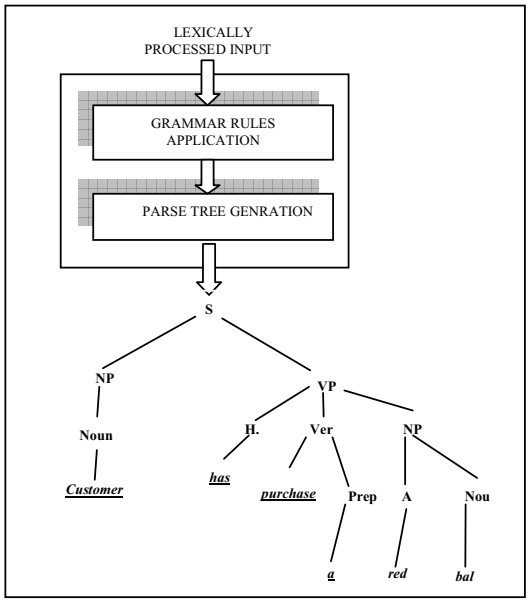


Figure 4.Syntactic Processing

C. *Anaphora Resolution*

In linguistics, anaphora is an instance of an expression referring to another. The process of anaphora resolution is the problem of resolving what a pronoun or a noun phrase refers to that is previously defined. In other words, the referential entity is called as an anaphor and the entity to which it refers is called as an antecedent .The process of determining antecedent of anaphor is called Anaphora Resolution. Algorithm for anaphora resolution is given in Fig. 5.

```
Algorithm: Anaphora Resolution
Input: Array of individual sentences (pre-processed)
Output: Resolved Anaphora
Procedure
begin
do
{
  //Anaphora resolution for he/she/it kind of word
    If
    there exist a word     Qi     in     sentence     Sj     ; such that
Pronoun(Qi) is true
    then
        if   there exists word  Qk  in  sentence   Sj-1  ;  such  that
Noun(Qk) is true
        then       Qi is anaphor of Qk
    If  //Anaphora resolution for  who/where kind of words
     there exist a word     Qi     in     sentence     Sj     ; such that
POS(Qi)="WH" is true
    then
       if  there exists word Qk in sentence  Sj ; such that Noun(Qk)
is true
          and Gender (Qi)=Gender(Qk)
          then       Qi is anaphor of Qk
}
 while(end of array)
```

Figure 5. Algorithm for Anaphora Resolution

The decision tree for anaphora resolution of *he/she/it* kind of words is shown in Fig. 6.

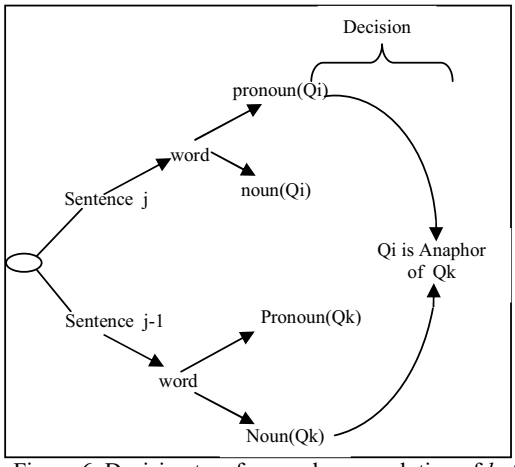


Figure 6. Decision tree for anaphora resolution of *he/she/it* kind of words

The decision tree for anaphora resolution of *who/where* kind of words is shown in Fig. 7.

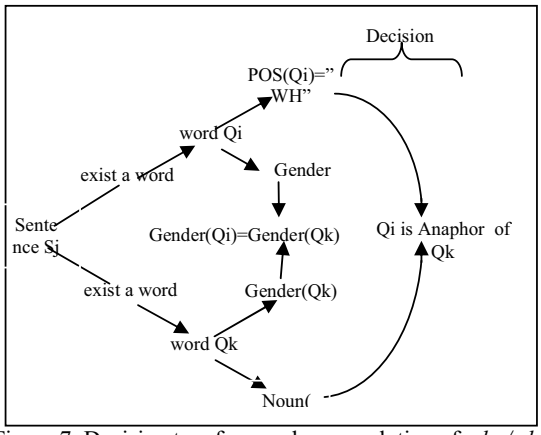


Figure 7. Decision tree for anaphora resolution of *who/where* kind of words

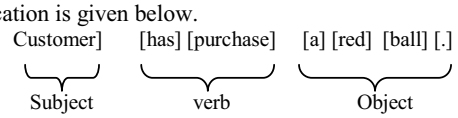
D. *Semantic Processing*

This phase helps in identifying the main parts of a sentence i.e. object, subject, actions and attributes. In this phase, associations are identified by doing semantic analysis. Which actions have been performed by which object and a set of attributes belong to which object is also determined in this phase. This phase comprises following sub phases:

1. Main parts identification
2. Thematic object and its attributes extraction

1) *Main Parts Identification*

In This step helps in identifying the main parts of a sentence i.e. object and subject. The example for main parts identification is given below.



2) *Thematic Object and its Attributes Extraction*

This module is responsible for extracting the thematic object and its attributes from the sentence. Any word that appears in

object part and is not followed by any other noun is called a thematic object. Adjectives in the sentence are specified as attributes of the thematic object. The example of thematic object and its attribute is shown below.

[Customer] [has] [purchase] [a] [red] [ball] [.]

Attribute Thematic object

Object nouns are sometimes specified as objects and sometimes as attributes. The algorithm for thematic object and its attributes extraction module is given in Fig. 8.

```
Algorithm: Them_object_attri_extraction
Input: Sentence with main part labelled
Output: Thematic object and its attributes
Procedure
begin
for each sentence
{   //identification of Thematic Object
    If there exist a word Q in sentence; such that Q appears in object part
    of sentence and not followed by any other noun.
     then : Q is a thematic object
  //identification of attributes
   if      there exist a word Q in the sentence S such that Q appears in
object
          part of sentence and Q is adjective of thematic object
   then :Q is attribute of thematic object
    if    there exist a word Q in the sentence S such that Q appears in
object
          part of sentence and Q is prefixed by  's   or Q is post fixed by
of
   then :Q is attribute of thematic object
  if       there exist a word Q in the sentence S such that Q appears in
object
         part of sentence and Q is post fixed by punctuation  ',' or 'and'
and
        followed by thematic object
  then :Q is attribute of thematic object
}end
```

Figure 8. Algorithm for thematic object extraction module

A decision tree for extracting the thematic objects and attributes is shown in Fig. 9.

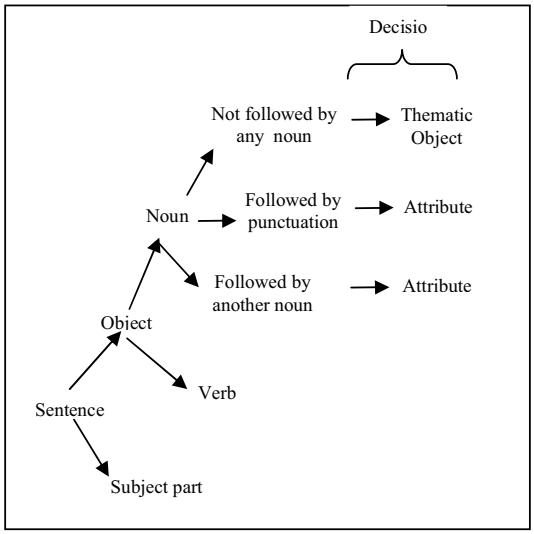


Figure 9. Decision Tree

E. Ontology Population

To provide valuable, consistent ontology based knowledge services, ontologies must be populated with many high-quality instantiations. Ontology population refers to the insertion of information into the knowledge base following the ontology domain representation. Manual ontology population is labour intensive and time consuming. A fully automatic feeding of Web-based knowledge to the ontology is presented here.

XML files, one per document, represent information extracted in ASDP with respect to a given ontology using tags mapped directly from Thematic objects and Attributes names (See Fig. 10).

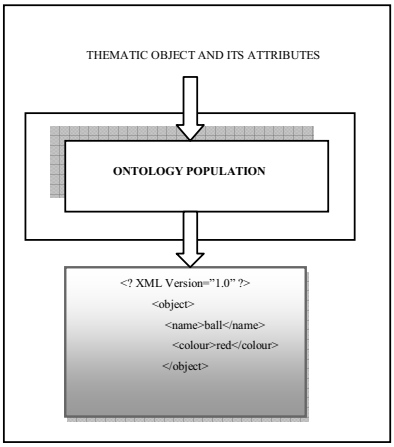


Figure 10. Ontology Population

V EXAMPLE

In the given example a sentence *"Prashant has purchased a hatchback ,sandriftgray, chevrolet, car. It has power-steering, power-window."* is supplied to the proposed system. The intermediate results and outcome are shown in Fig. 11(a) and Fig. 11(b).

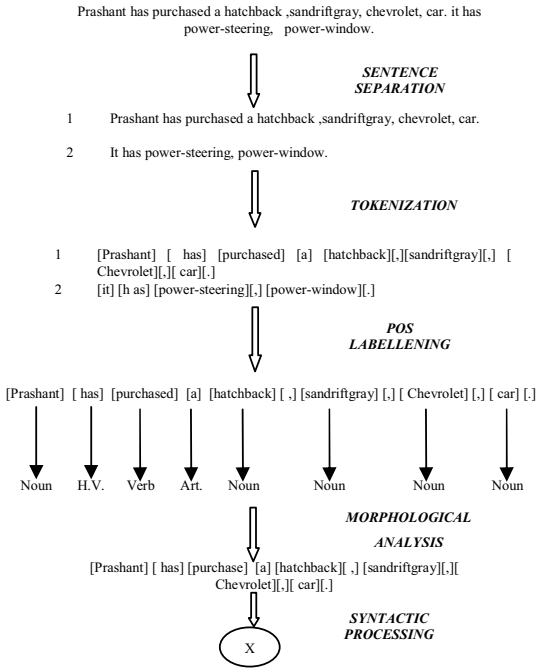


Figure 11(a). Example -A

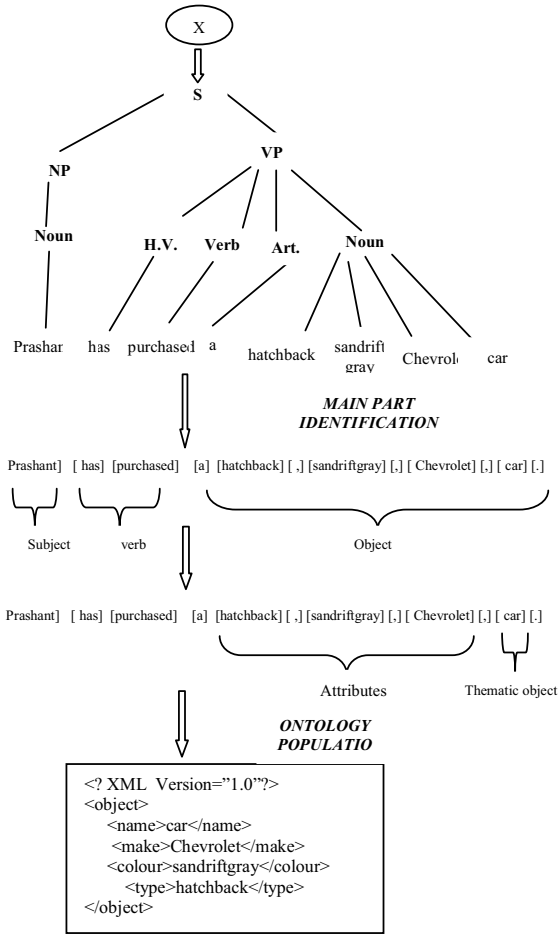


Figure 11(b). Example -B

VI CONCLUSION

In this work, a technique of Automatic Semantic Domain-ontology Populator (ASDP) towards construction of given domain modelled by the database is proposed. ASDP is a way to find, structure, and display relationships between concepts, which consist of attributes and objects. This method helps in understanding a given domain and in building a domain model for it.

In the present work, It has been identified that the five phases are required for the automatically extracting the semantic information from the text documents and populate them in ontology.

1. In the first phase Lexical Processing of a plain text is performed.
2. In the second phase English grammar rules are applied to the input received from lexical processing phase for performing Syntactic Processing.
3. The process of Anaphora Resolution is the problem of resolving what a pronoun or a noun phrase refers to that is previously defined in the document is performed in next phase.
4. In Semantic Processing phase main parts of a sentence i.e. object, subject, actions, attributes are identified.
5. Finally in Ontology Population phase, the information is inserted into the Knowledge Base following the ontology domain representation.

VII FUTURE WORK

The future work is to address the problem of resolving cataphora (twin domain of anaphora) where the occurrence of the pronoun is to be at an initial stage and noun shall be occurring later in the sentence. The future work will also try to identify the various relationships among the thematic objects such as aggregation, association, generalization, specialization and equivalence.

REFERENCES

[1] Brigitte Biebow and Sylvie Szulman. Terminae: A linguistic-based tool for the building of a domain ontology. In Knowledge Acquisition, Modeling and Management, pages 49–66, 1999.

[2] Deryle Lonsdale, Yihong Ding, David W. Embley, and Alan Melby. Peppering knowledge sources with salt: Boosting conceptual content for ontology generation. In AAAI Workshop for Semantic Web Meets Language Resources, The Eighteenth National Conference on Artificial Intelligence, pages 30–36. AAAI Press, 2002.

[3] DA-YOU LIU. Learning owl ontologies from free texts. In Machine Learning and Cybernetics, volume 2, pages 1233 – 1237, 2004.

[4]Latifur Khan and Feng Luo. Ontology construction for information selection. In Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence, ICTAI '02, pages 122–, Washington, DC, USA, 2002. IEEE Computer Society.

[5] Rob Engels, Dieter Fensel, Frank van Harmelen, Victor Iosif, Arjohn Kampman, Uwe Krohn, Ulrich Reimer, Rudi Studer and York Sure Content-driven Knowledge Management through Evolving Ontologies On-To-Knowledge IST-1999-10132.

[6] Mehrnoush Shamsfard, Ahmad Abdollahzadeh Barforoush **:**Learning ontologies from natural language texts. International Journal of Human-Computer Studies Volume 60 , Issue 1 (January 004) Pages: 17 – 63 Year of Publication: 2004 ISSN:1071-5819 Academic Press, Inc. Duluth, MN, USA

[7] Michel Klein, Dieter Fensel, Frank van Harmelen, and Ian Horrocks. The relationbetween ontologies and xml schemas. Electronic Transactions on ArtificialIntelligence, 2001.