International Journal of Emerging Trend in Engineering and Basic Sciences (IJEEBS)
*ISSN (Online)*2349-6967 *Volume 2 , Issue 1(Jan-Feb 2015), PP1-5*

Semi-Automated Building Ontology's from Unstructured Text

Varsha Gadhave[1], Shweta Benke[2],Shrikrishna Jadhav[3],Sheetal Khobragade[4]
Prof. Yogesh Sayaji[5]
Information Technology Dr.D.Y.Patil College of engg.University Of Pune, India

Abstract: -Ontology is playing an increasingly important role in knowledge management and the semantic web. Many applications call for methods to enable automatic extraction of structured information from unstructured natural language text. Due to the inherent challenges of natural language processing, most of the existing methods for information extraction from text tend to be domain septic. It is also an efficient and extensible text mining system that could be used in many applications related to natural language document processing. Many applications call for methods to enable automatic extraction of structured information from unstructured natural language text. Due to the inherent challenges of natural language processing, most of the existing methods for information extraction from text tend to be domain septic. This thesis explores a modular ontology-based approach to information extraction that decouples domain-specific knowledge from the rules used for information extraction.

Keywords: -Natural language processing, Ontology, Text mining, Automatic Question Generation

I. INTRODUCTION

Ontologies are an importance in many application areas where the intelligent knowledge base integration and natural–language processing. Their importance is growing, as is growing on the Web the number of information repositories that need metadata enrichment and analysis. Whereas an ontology engineering which is very time consuming and expensive and growing need for automated or at least semi-automated Each key phrase is matched with a Wikipedia article and classified into one of five abstract concept categories: Research Field, Technology, System, Term, and Other. One more thing which is is been added in this paper particularly for the better and advance feature is an Automatic Question Generation (AQG) is a challenging task which involves natural language understanding processing and generation We propose creating a semi-automated system that generates concept maps to easily manage knowledge bases. Concept maps tend to make the structure of a body of knowledge much more significant for human users than other forms of knowledge representation. Hence, they are more easily validated and enriched by a domain expert. Concept maps also foster meaningful learning and index sentences at a one-grained level, which is required for indexing and retrieval. It's an true in domain of online training of online training. (ITS)Intelligent Tutoring System is the resources to build a Knowledge base

1.1 Construction of Conceptual Graph

As key phrases are classified then after next step conceptual graph formation takes place based on section content and headings on the Wikipedia article. A conceptual graph includes nodes representing key phrases and sub phrases. Graph may containdefinitions, drawback, advantages, subtypes, limitations etc.

1.2 Wikipedia as Background Knowledge Source

We can see Wikipedia as a lexical semantic resource which contain information about named entity and domain specific terms.In manynatural language processing tasks it has been successfully applied.Wikipedia is used for reasons like – it has more than three million pages, have large knowledge base, it covers multiple domains too.These Wikipedia articles are used to construct conceptual map and graph structures. Java Wikipedia Library (JWPL) is open-source and Java based API's that we used to overcome problems withWikipedia in XML dumps as they are not programmatically accessible.It parses Wikipedia articles with WML (Wikipedia markup language) and converts it into databases.

1.3 Automatic Question Generation System

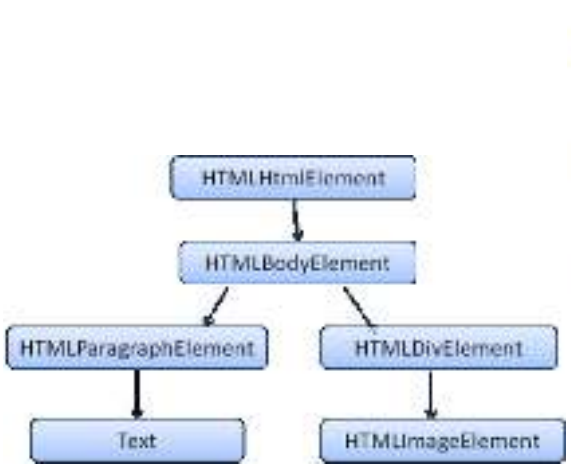
AQG is approach which is based on pattern matching rules and it contains transformation of declarative sentence into a question. The questions generated are of type why, how, what. Some of other approaches are based on automatic multiple choice QG. The distractors like hypernyms and hyponyms of the term were identified by referring WordNet.Mitkov and Ha revealed that automatic question generation and manual correction is time efficient than manual question generation. However, it is not in providing feedback in current project rather it is focused on generating assessment items for the base concept.

The current study based on previous studies. In this study, we are proposing a novel approach to address key challenges of automatic trigger question generation. The first challenge is identification of key or a central concept from the base concepts. The another is related with the systems lack of knowledge about the domain. Another main approach is the question generated is useful, helpful to user or not.

1.4 Key Phrase Extraction Technique

Key phrases provide the important information about concept. For the extraction of key phrases system uses an unsupervised extraction algorithm. System classifies and checks each key phrase with the Wikipedia article using rule-based approach. The key phrases belongs to one of the Research Field,Technology, System,Term , Other. A system called GenEx which is developed by Turney for the extraction of key phrases.For improving results Naı¨ve Bayes classifier for the extraction of key phrases applied by Frank et al. Both Naı¨ve Bayes classifier and GenEx are examples of supervised approaches to key phrase extraction. Clustered terms are the terms which share the similar noun terms from a list of extracted noun phrases.

The Lingo algorithm is used for clustering web search results which is based on based on singular value decomposition. In Lingo algorithm from input documents term-document matrix A is built. This matrix is broken intothe three matrix(U,S,V) such that A= USVT.



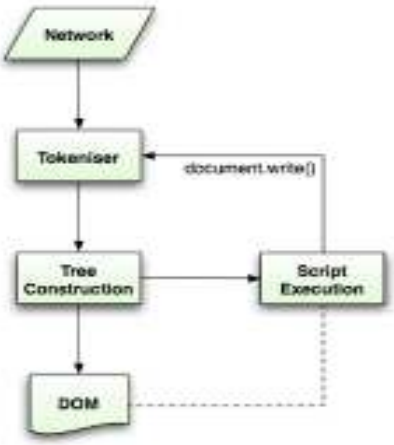


Fig 1.1 Example tree creationFig 1.2 HTML parsing

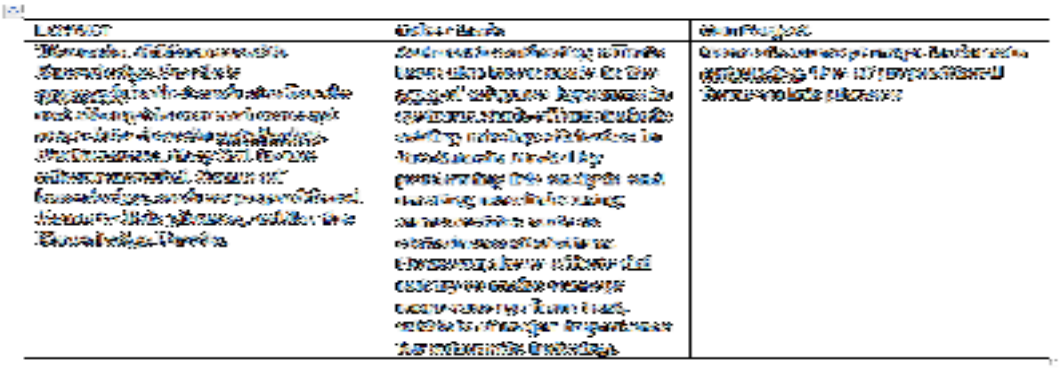
II. LITERATURE SURVEY

1)T.Zesch, C. Mu ller, and I. Gurevych, Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary, Proc. Sixth Intl Conf. Language Resources and Evaluation, 2008. DYPCOE Department o9f Information Technology, Pune 2014.

2)M. Heilman and N.A. Smith, Extracting Simplified Statements for Factual Question Generation, Proc. Third Workshop Question Generation, 2010.

3)A.M. Olney, W.L. Cade, and C. Williams, Generating Concept Map Exercises from Textbook, Proc. ACL HLT Workshop Innovative Use of NLP for Building Educational Applications, 2011.

3.1 Comparative Analysis



III. MATHMATICAL MODULE

3.1 Algorithms

- Web scraping to extract text data from html

- NLTK grammar type for each word extraction e.g. nouns, verbs , etc.
- Anaphora resolution to convert pronouns into nouns.
- Simple matching to identify key phrases and relationships e.g.Inheritance is part of oops. Thus oops is the root node and inheritancecomes under that node.

A. Web Scrapping Algorithm:

```
function webscrape() {
contents = read_content(url)
 parse_html(contents)
}

function parse_html(contents) {
  tokens = tokenize(contents)
}

function tokenize(contents) {
 do {
   node.starttag = identify_start_tag()
node.attributes = identify_attributes()
  node.text = identify_text()
  node.endtag = idenify_end_tag()
 node.parent = identify_parent()
  insert_into_tree(node)
 } while(node.endtag != '</html>')
}
```

B.Anaphora Resolution:

Anaphora resolution, NLTK breaking of words using existing python / java libraries e.g. Stanford Parser, Rap anaphora, python nltk.

3.2 Comparative Evaluation

This approach is based on set of metrics that consider the ontology a graph entity.it consist of Class Match
Measure (CMM), the Between-ness Measure (BEM), the Density Measure (DEM), and, finally, the Semantic Similarity
Measure (SSM).For ranking ontologies with respect to terms sought the total score of these four is added.

3.2.1 The Class Match Measure(CMM)

It evaluates the coverage of an ontology for the provided keywords . Given the input keywords, the ONTO-EVALUATOR searches through the ontology classes to determine if the keywords are expressed as classes (exact match) or if they are included in class labels (partial match).

3.2.2The Density Measure (DEM)

The DEM expresses the degree of detail or the richness of the attributes of a given concept. It is assumed that a satisfactory representation of a concept must provide sufficient detail regarding its nature.

3.2.3The Between-ness Measure (BEM)

The BEM calculates the between-ness value for each sought term in the generated ontologies. It measures the extent to which a concept lies on the paths between others. Class centrality is considered important in ontologies.

3.2.4 The Semantic Similarity Measure (SSM)

The last measure, the SSM, computes the proximity of the classes that match the sought keywords in the ontology. As Alani and Brewster stated, if the sought terms are representative of the domain, the corresponding domain ontology should link them through relationships (taxonomic or object properties). Failure to do so may indicate a lack of cohesion in the representation of the domain knowledge.

Finally, based on these four metrics, an overall score is computed. Let M={M[1],M[2],M[3],M[4]}={CMM,DEM, SSM,BEM} wi be a weight factor, and O be the set of ontologies to rank. The score is computed as follows [2]:

computed. Let M = {M[1], M[2], M[3], M[4]} = {CMM, DEM, SSM, BEM}, wi be a weight factor, and O be the set of ontologies to rank. The score is computed as follows [2]:

$$Score(o \in O) = \sum_{i=1}^{4} w_i \frac{M[i]}{\max_{1 \le j \le |O|} M[j]}.$$

User enters search text. we have to show the user the right ontology. Thisis done using 4 scores. Class match measure, Density measure, betweenmeasures, semantic similarity measure.

IV. EXISTING SYSTEM

- Every company has large amount of unstructured documentation relating to a domain e.g. Banking, Insurance etc.
- It is very expensive to maintain this knowledge base
- It is a time consuming process for the new comers to read & learn from the huge knowledge base.

4.1 Problem Definition

Given unstructured text, generate concept map. Also based on the concept maps generate questions based on singleconceptual graph structure. We have adopted two principles toguide the design of our question templates. First, the questions should be specific. We place the description of a key concept inthe beginning. Second, the questions should be linked to theauthor's research. We place the judgmental questions after orcombined with the description of the key concept.

V. PROPOSED SYSTEM

5.1 System Description

Our system reads unstructured text. It then uses natural language processing to extract key phrases. It then generates concept maps and questions.

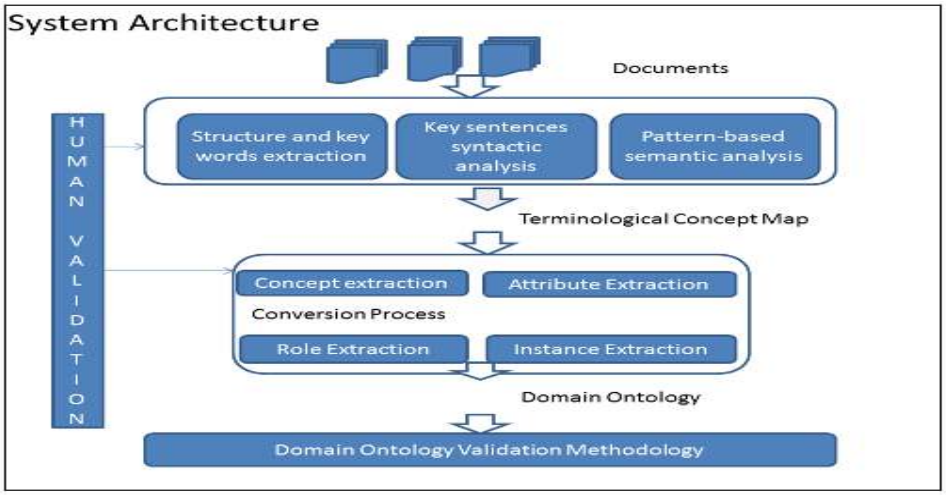


Fig 5.1 System Architecture

5.2 System Workflow

Key Phrase Extraction In the preprocessing stage, all input documents (literature review papers) are split into sentences. A term-sentence Vector Space Model (VSM) is then built. In stage 1, the keyphrase extraction based on the VSM was performed using the Lingo algorithm. The key phrases extracted can be used in different ways. A fully automatic system could use the key phrases unaltered or use a blacklist prepared once (not in runtime) by a domain expert, and reused as needed. If the system is not used regularly, a blacklist can be made ad hoc (as is the case in this study).If the extracted key phrase is in abbreviated form(acronym), its full name was searched by using regular expression pattern matchingtechniques to increase the chances of bonding matching Wikipedia articles.

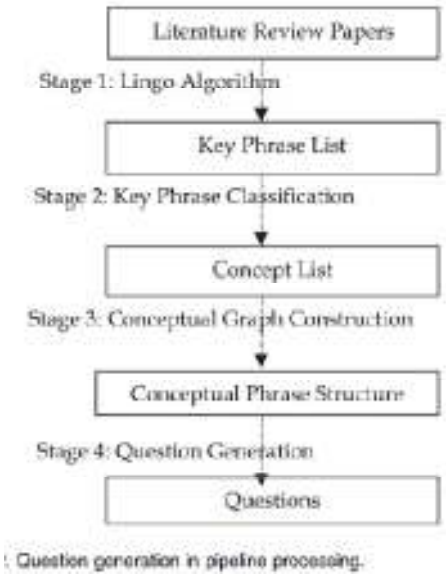


Fig. 5.2 Work Flow Diagram

VI. CONCLUSION

Generally any human being will require ample amount of time for understanding any particular concept. Any learner or any person may get Concept map for better and flexible understanding of any particular concept using semi-automated ontology building system. So overcoming this problem concept map is an easy technique.However, traditional assessments are mostly measures the objectivity of the learning, and evaluate the learner as achievement in the form of marks. Nevertheless, concept map assesses learner as subjectivity and it interprets in the form of marking and grading. Concept map is a prominent assessment tool and its scoring technique assesses the actual knowledge structure of thestudents. For that, expert concept map is the standard, it assesses the maximum number of concepts, links, and propositions, and it helps the student as map to count number of propositions and concepts at their level. In this assessment technique, the score represent the student as actual understanding and their subsequent knowledge structure. Some evidence from other researchers, Therefore, the present study has come under the Novakian area of search and the finding also importance to the world of education. Hence, the cooperative (collaborative) learning an important attribute in the curriculum so as to educating students for coping in today as world.

1. We are using all open source tools.
2. Thus we do not have any economic cost associated with the project apart from our efforts.
3. We automatically create the structured text from unstructured text.
4. This reduces the readers workload and time.

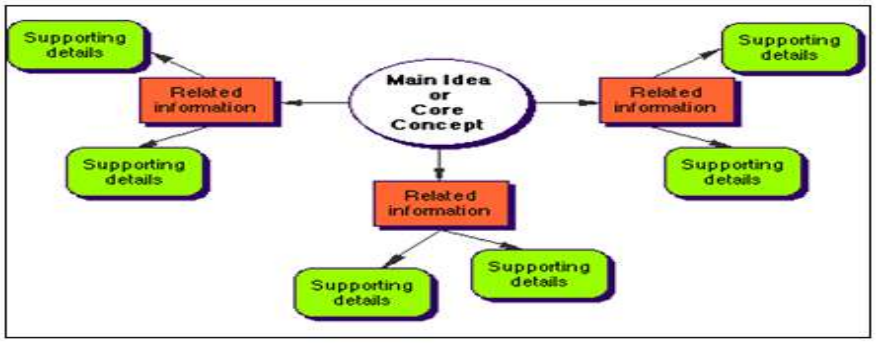


Fig. 6.1Output Structure

VII. ACKNOWLEDGEMENTS

Every engineering student looks toward the final year project as an opportunity by which he can implement the skill that he has eventually nurtured in the year by hard work dedication the milestone of completing the project would have been intractable without the help of few people who need to be acknowledge.
We owe this moment of satisfactions with a dear sense gratitude to our internal guide Prof.Yogesh Sayaji who guided us at every stage. Whose technical support and helpful attitude give us high moral support. We would also like to extend our sincere thanks to our H.O.D. Prof. Ravi Patki for his guidance and constant encouragement.
We are highly obliged to the entire staff of the information technology department and principal sir for their kind co-operation and help. We also take this opportunity to thank all our colleagues who baked our interest by giving useful suggestions and also possible help. At last but not least we are thankful to our friend colleagues and all the people directly or indirectly concerned with this project

REFERENCES

[1] G. Graswell, Writing for Academic Success: A Postgraduate Guide. SAGE, 2008.

[2] B. Steward, ^aWriting a Literature Review,^a The British J. Occupational Therapy, vol. 67, pp. 495-500, 2004.

[3] M. Afolabi, ^aThe Review of Related Literature in Research,^a Int^al J. Information and Library Research, vol. 4, pp. 59-66, 1992.

[4] D. Taylor, ^aThe Literature Review: A Few Tips on Conducting It,^a http://www.writing.utoronto.ca/advice/speci_ctypes- ofwriting/ literature-review, July 2011.

[5] V. Rus and A.C. Graesser, ^aThe Question Generation Shared Task and Evaluation Challenge,^a Proc. Sixth Int^al Natural Language Generation Conf., pp. 251-257, 2009.