ISSN NO. 2320-5407

Journal homepage:http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

National Symposium On Emerging Trends In Computing & Informatics, NSETCI 2016, 12th July 2016,Rajagiri School of Engineering & Technology, Cochin, India

An Automated Mapping Process Using NLP Technique To Correlate Program Objectives And Outcomes Offered By Educational Institutions.

***Jithin S Kuruvila, Midhun Lal, Rejin Roy and Sandy Joseph.**
Department of Computer Science and Engineering,Rajagiri School of Engineering, India.

Manuscript Info *Abstract*

Key words:
Automated Mapping,
NLP Technique,
Stemming,
Lemmatization,
Semantic Analysis

In this technology evolving era expansion of higher education sector is essential and is getting multiplies every year. Accreditation is the principal means of providing society with reliable information on quality of education offered by the institutions along with the expansion. The National Board of Accreditation (NBA) is a quality assurance scheme for higher technical education and conducts evaluation of programs offered by technical institutions on the basis of various criteria include but not limited to Vision, Mission and Programme Educational Objectives (PEOs), Programme Outcomes (POs), Courseoutcomes (COs) etc.. The correlation (mapping) between the POs and the COs is to be provided in order to establish the contribution of the POs towards the attainment of the PEOs. The mapping process facilitates the institute to judge the overall effectiveness of its own processes. Mapping process requires an unlimited amount of human effort, time and patience. To reduce human task and time of mapping this paper proposes an automated mapping technique using natural language processing technique. To ensure fast and efficient mapping this approach adopts Bloom's taxonomy based lemmatization technique and wordnet based semantic analysis that facilitates the institute to judge the overall effectiveness of its own processes.

Introduction:-

In an increasingly technologically dependent world, expansion of higher education sector is imperative in an emerging economy. Hence the number of institutions as well as programs offered has multiplied. The challenge is to ensure its quality to the stakeholders along with the expansion. Accreditation is the principal means of quality assurance in higher education. It provides society with reliable information on quality of education offered by the institutions and helps the Institution to know its strengths, weaknesses and opportunities. In achieving recognition, the institution or program of study is committed to meet certain minimum specified standards and also seeks ways to enhance the quality of education. The National Board of Accreditation (NBA) is a quality assurance scheme for higher technical education and conducts evaluation of programs offered at various levels by technical institutions on the basis of laid down norms. But, accreditation provides quality assurance that the academic aims and objectives of the Institution are honestly pursued and effectively achieved by the available resources and that the Institution has demonstrated capabilities of ensuring effectiveness of the educational programme(s). Assessment and accreditation shall be based on various criteria. This may include but not limited to Vision, Mission, Programme Educational Objectives (PEOs), Programme Outcomes (POs) etc.

The Program Educational Objectives are broad statements that describe the career and professional accomplishments that the programme is preparing graduates to achieve. Programme Outcomes are narrower statements that describe what students are expected to know and be able to do upon the graduation. These relate to the skills, knowledge, and behaviour that students acquire in their matriculation through the programme. Course Outcomes define the qualities attained by the students on completing the particular course on a subject.

The POs formulated for each programme by the institute must be consistent with the NBA's Graduate Attributes. The POs must foster the attainment of the PEOs. The correlation between the POs and the COs is to be provided as per the format given in order to establish the contribution of the POs towards the attainment of the PEOs. NBA adopts a mapping technique to evaluate the achievement of programme educational objectives and programme outcomes. Mapping may be done for one to many, many to one, and many to many parameters. This process facilitates the institute to judge the overall effectiveness of its own processes. Mapping process requires an unlimited amount of human effort, time and patience, an automated mapping technique can reduce human task and time. This mapping helps the institution to ensure that the programme curriculum that was developed at the time of inception of the programme has been refined in the subsequent years to make it consistent with the PEOs and the POs.

This paper is organized as follows. A brief review of some of the existing document comparison techniques the basic ideas related to the work are given in section 2. The proposed system architecture is described in section 3. Section 4 gives the implementation details and test results and Section 5 concludes the paper.

Related work:-

In this digital age it is quite important to retrieve information from natural language. The major processes in natural language processing (NLP) are tokenization, Morphology and morphological processing, Syntax and syntactical processing, Semantics and semantic processing

Tokenization is a process of identification of token/topics within input documents/statements and helps to retrieve precise information with reduced search(S. Raman et.al, 2012) and storage space (S. K. M. Wong et. al, 1985) Tokenization involves pre-processing of documents and generates its respective tokens (Vikram Singh and Balwinder Saini, 2014). Various traditional techniques are designed for tokenizations, Porter's algorithm (M.F.Porter,1980) is one of the best tokenization techniques, but this algorithm suffers from accuracies during the identification and efficiency (Poroshin V.A, 2004). After splitting the words into meaningful tokens morphological processing is done to identify morphological classes these words belong to. One of the widespread tasks is lemmatizing or stemming, in this case all morphological variations of a given word are collapsed to one lemma or stem. Stemming is done to make sure that variants of words are not left out when texts are retrieved. Stemming techniques are many, including the Paice/Husk stemmer (D. A. Hull, 1996), Porter's stemmer and Lovin's stemmer (J. B. Lovins, 1968). In the Paice/Husk stemmer, a file is created which holds a set of rules, and these rules are read by an array which implements the rules until a final stem is achieved. It accepts and processes a rule if the word specifies an ending which matches the last letters of the word. The Lovins stemmer is a single pass, context-sensitive algorithm which only removes one suffix from a word by utilizing a list of 250 suffixes and removing the longest suffix that it finds attached to the given word. Lemmatization on the other hand uses vocabulary and morphological analysis of word and tries to remove inflectional endings, thereby returning words to their dictionary form. It checks to make sure that things are done properly by analyzing if query words are used as verbs or nouns. Lemmatization also helps to match synonyms by the use of a thesaurus. In syntactical processing a linear sequence of tokens are transformed into a hierarchical syntax tree. In this level part of speech tagging, chunking or detecting syntactic categories (verb, noun phrases) and sentence assembling (constructing syntax tree) are handled.In the fourth phase Semantics analysisfinds synonyms, word sense disambiguation and populate base of knowledge.

There are a few techniques have been developed by researchers for plagiarism detection in natural language documents. One of the common techniques is fingerprint matching technique. Comparing two documents under this technique is counting the number of substrings that are common in both fingerprints (Heintze andNevin, 1996). In 2003, Lyon et al. generates fingerprint using phrase mechanism to measure the resemblance between two documents. This approach converts each document to a set of trigrams (three words). Then, the set of trigrams for each document is compared with all other using the matching algorithm. Finally, the measure of the resemblance for each pair of documents is calculated. The pros and cons of character-based and phrase-based finger printing have led (Yerra and Ng 2005) to represent the fingerprints of each statement by three least-frequent 4-grams. The processing

time and space usage for this method are much better than the character-based and phrase-based techniques. However, in the case of rewording and restructuring of statements, all fingerprinting techniques will fail in detecting the similarity. (Chowet.al,2010) integrate the fingerprint matching technique with a feature based text similarity detection approach.

Proposed Approach:-

Our proposed system automates the mapping procedure of the course objectives and outcome with the program objectives and outcomes that are been set according to the NBA norms. This proposed mapping process could greatly reduce the amount of time and human efforts required for the mapping procedure. In this approachthe two comparing statements are tokenized in to words followed by the removal of stop words. To tokenize the statements, text needs to be segmented into linguistic units such as words, punctuation, numbers, alpha-numeric, etc. The next step is to remove all these stopwords and punctuations. Next step is to do the stemming (removing suffixes by automatic means). After stemming process lemmatization is also done to convert these words to the exact matching words as given by the Bloom's, before doing the string comparison to determine the taxonomy levels. Next process is to determine the levels of the separated action verbs in the two statements. The Python data structure "list" is mainly used in this module to store the Bloom's Taxonomy action verbs at the six levels that are shown in the fig. 1. Then after the stemming and lemmatization process the word by word semantic analysis is performed with the help of the data structure called WordNet. Since this comparison is done on the technical program objectives and outcome lot of technical words that are not available in the Wordnet list are come into picture. Thus to improve the efficiency of the mapping we have also created a technical word list pertain to a specific stream of study. If the number of similar words exceeds the threshold limit the two sentences can be mapped or else the statements are not comparable. To reduce the comparison the act ion verbs are classified according to Bloom's Taxonomy and verbs at the same level or higher are only need to be compared.The proposed system architecture is shown in Fig. 1(a) and the process diagram is shown in Fig. 1(b).

Bloom's Taxonomy was created by Benjamin Bloom during the 1950s and is a way to categorize the levels of reasoning skills required in classroom situations. Bloom identified six levels of cognitive complexity that have been used over the past four decades to make sure that instruction stimulates and develops students' higher-order thinking skills. To ensure quality in higher education, while setting the objectives and outcomes for the programme and courses, subject experts incorporate the various taxonomy levels. The actions words for various six levels of Bloom's Taxonomy: Knowledge, Understand, Apply, Analyze, Evaluate, Create are shown in Fig. 2.

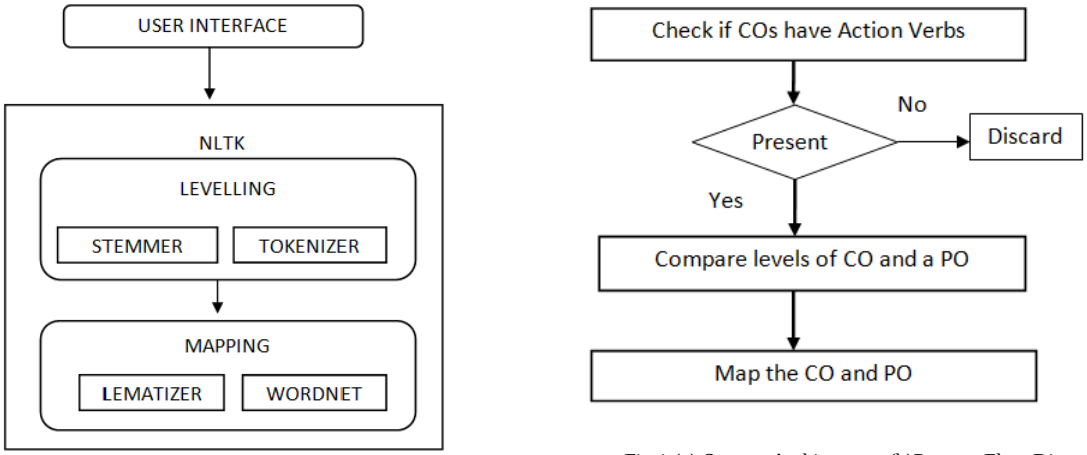

Fig.1 (a) System Architecture; (b)Process Flow Diagram

Action Words for Bloom's Taxonomy					
Knowledge	**Understand**	**Apply**	**Analyze**	**Evaluate**	**Create**
define	explain	solve	analyze	reframe	design
identify	describe	apply	compare	criticize	compose
describe	interpret	illustrate	classify	evaluate	create
label	paraphrase	modify	contrast	order	plan
list	summarize	use	distinguish	appraise	combine
name	classify	calculate	infer	judge	formulate
state	compare	change	separate	support	invent
match	differentiate	choose	explain	compare	hypothesize
recognize	discuss	demonstrate	select	decide	substitute
select	distinguish	discover	categorize	discriminate	write
examine	extend	experiment	connect	recommend	compile
locate	predict	relate	differentiate	summarize	construct
memorize	associate	show	discriminate	assess	develop
quote	contrast	sketch	divide	choose	generalize
recall	convert	complete	order	convince	integrate
reproduce	demonstrate	construct	point out	defend	modify
tabulate	estimate	dramatize	prioritize	estimate	organize
tell	express	interpret	subdivide	find errors	prepare
copy	identify	manipulate	survey	grade	produce
discover	indicate	paint	advertise	measure	rearrange
duplicate	infer	prepare	appraise	predict	rewrite
enumerate	relate	produce	break down	rank	role-play
listen	restate	report	calculate	score	adapt
observe	select	teach	conclude	select	anticipate
omit	translate	act	correlate	test	arrange
read	ask	administer	criticize	argue	assemble
recite	cite	articulate	deduce	conclude	choose
record	discover	chart	devise	consider	collaborate
repeat	generalize	collect	diagram	critique	collect
retell	give examples	compute	dissect	debate	devise
visualize	group	determine	estimate	distinguish	express
	illustrate	develop	evaluate	editorialize	facilitate
	judge	employ	experiment	justify	imagine
	observe	establish	focus	persuade	infer
	order	examine	illustrate	rate	intervene
	report	explain	organize	weigh	justify
	represent	interview	outline		make
	research	judge	plan		manage
	review	list	question		negotiate
	rewrite	operate	test		originate
	show	practice			propose
	trace	predict			reorganize
	transform	record			report
		schedule			revise
		simulate			schematize
		transfer			simulate
		write			solve
					speculate
					structure
					support
					test
					validate

Experimental Setup and Result Analysis-

This software is developed and tested using Python 2.7 programming language in Ubuntu 12.04 OS on core i5, 1.7 GHz processor with 4GB RAM. The proposed software system will feed the inputs as the POs and COs of various programs and courses. This approach has two phases. In the first phase action verbs of the COs and POs of various subjects are identified after the tokenization and stemming process, then determines the educational levels of the various action verbs. If any CO does not meet the criteria it is filtered and screened for the ease of comparison. Pseudo code of phase I is given below.

```
Level( )
        while( end of file ):
                read the next CO
                check for action verbs
                if not found then:
                        discard the CO
                else:
                                add the CO to the list in the level in which the action verb belong
                end if
        end while
        add the lists to the file Level.txt
    repeat the same steps for PO and store the list
end
```

The mapping process takes place in the second phase. Mapping of COs and POs are done based on the word by word semantic analysis of the two statements with the help of the data structure called WordNet. If the number of similar words exceeds the threshold limit the two sentences can be mapped or else the statements are not comparable. The main component used that is used in this module is the wordnet. It is a module that is used to compare the semantic meaning between two words. The output of this comparison will be a score which shows the degree of similarity between the two words.

In this module, the input is taken from the file which has the COs and POs that are tokenized and arranged based on the Bloom's Levels. Then each CO is compared to each PO word by word using wordnet and generates an average score. The PO-CO combination that produced the highest score will be mapped. Sample PO and CO are given in section 4.1 and 4.2 and the mapping result is shown in Fig. 3.Experimental results show that proposed software is capable to efficiently generate mapping score with less computational and time complexities.
Pseudo code for the mapping module is shown below:

```
Map( )
        while( not end of CO file ):
                read the next CO
                while( not end of PO file ):
                        read the next PO
                compare the CO to the PO
                        generate the average score
        end while
                map the pair with highest average score
        end while
        print the result as a table
end
```

Results and Discussion:-

Sample Program Outcomes (POs):-

1) Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2) Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusion using first principle of mathematics, natural science, and engineering science.
3) Design/development of solution: Design solution for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, social, and environmental considerations.
4) Conduct investigations of complex problem: Uses research-based knowledge and research method including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusion.
5) Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.
6) The engineering and society: Apply reasoning informed by the contextual knowledge to asses societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7) Environment and sustainability: Understand the impact of professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
8) Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9) Individual and team work: Function effectively as an individual, and as a member or leader in diverse team, and in multidisciplinary settings.
10) Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
11) Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
12) Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Sample Course outcomes COs):-

1. Students will be able to understand and **define** the various algorithm analysis techniques.
2. Students will be able to **differentiate** and **give examples** for the different types of algorithm development strategies.
3. Students will be able to **estimate** the asymptotic time complexities of various recurrence relations
4. Students will be able to **select** the appropriate method to solve the recurrence relations
5. Students will be able to **use** Lower Bound Theory to improve the complexity of an algorithm

Since due to the time frame limit this work put effort in creating technical word list for computer steam only. Future enhancement of this work is to create a technical Wordnet that covers all technical stream so that all technical program evaluation can be done in a fully automated manner.

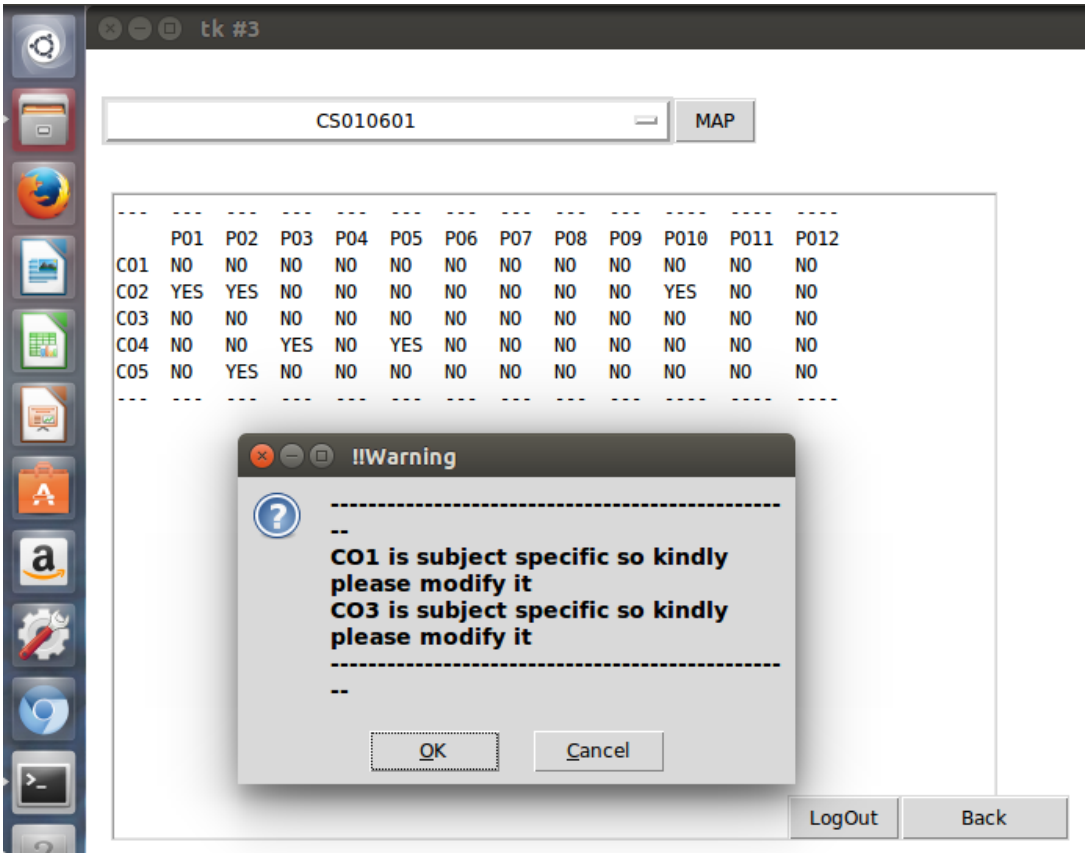

Fig. 3 POs and COs mapping Result.

Conclusion:-

In this technology dependent world the number of educational institutions is increasing every year. With this rate of increase, it is equally important to maintain the quality of such institutions. There are several methods by which the quality of an institution can be set, one of such methods is Accreditation. National Board for Accreditations (NBA), is an authorized body entrusted with the task of undertaking accreditation of technical education programmes. NBA adopts a mapping technique to evaluate the achievement of programme educational objectives and programme outcomes offered by the institutions But mapping is a time consuming process and requires lot of human effort. To reduce human effort and time of mapping this paper proposes an automated mapping technique using natural language processing. To ensure fast and efficient mapping this approach adopts Bloom's taxonomy based lemmatization technique and Wordnet based semantic analysis that facilitates the institute to judge the overall effectiveness of its own processes.

References:-

1. **S. Raman, V. Kumar, S. Venkatesan**, "Performance Comparison of Various Information Retrieval Models Used in Search Engines", IEEE Conference on Communication, Information and Computing Technology, Mumbai, India, 2012.
2. **S. K. M. Wong, W. Ziarko, P. C. N. Wong**, "Generalized vector space model in information retrieval," in the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, 1985, p. 18-25.
3. **Vikram Singh and Balwinder Saini. (2014).**An Effective Tokenization Algorithm For Information Retrieval Systems, CS & IT-CSCP, AIAPP – 2014, p. 109–119
4. **M.F.Porter**, (1980).An algorithm for suffix stripping,

5. **Poroshin V.A.**(2004) Semantic analysis of Natural Language, Chapter I. Basic Formalizing Principles Of Natural Language, p. 1-14
6. **D. A. Hull, (**1996),"Stemming algorithms: A case study for detailed evaluation," Journal of the American Society for Information Science, vol. 47, p. 70-84.
7. **J. B. Lovins**, (1968) "Development of a stemming algorithm," Mechanical Translation and Computational Linguistics, vol. 11, p. 22-31.
8. **Heintze and Nevin. (1996)** Scalable document fingerprinting. Proceedings of the Second USENIX Workshop on Electronic Commerce. Oakland, California.
9. **Lyon, C. Barrett, R. Malcolm, J.A**. (2003). Experiments in plagiarism detection. Technical report 388. School of Computer Science, University of Hertfordshire.
10. **Yerra and Ng. A. (2005)** Sentence-Based Copy Detection Approach for Web Documents. Fuzzy Systems and Knowledge Discovery.
11. **Chow Kok Kent, NaomieSalim**,(2010) Features Based Text Similarity Detection, Journal Of Computing, Vol. 2, no 1.