JOURNAL OF SOFTWARE ENGINEERING & INTELLIGENT SYSTEMS
ISSN 2518-8739
31st December 2017, Volume 2, Issue 3, JSEIS, CADMEI Copyright © 2016-2017
www.jseis.org

Intelligent classification of learning objects using information content, intra document terms and domain vocabulary

[1] Imran Ihsan, [2] Faisal Fayyaz Kiyani
[1,2]Department of Computer Science, Air University, Islamabad, Pakistan
Email: [1]iimranihsan@gmail.com, [2]faisalfk@gmail.com

ABSTRACT

LMS, databases of learning objects, are used by teachers to store, search, and retrieve learning objects. Classification of these learning objects is a tedious job. Metadata standards are available in order to specify a learning object; however, a taxonomic path is normally left for the developer of the application to decide. A common taxonomic path consists of various domains and sub-domains in the form of a hierarchy. Annotators decide to place a particular learning object in a specified domain but this is a time consuming and laborious work. Automatic and intelligent classification of these learning objects in their respective domain is a great challenge. Each learning object has a pedagogical content and that content can be measured by various techniques. In this paper, we will try to find Information Content in a learning object and classifying it using intra-terms co-occurrences and their frequencies. By using this inverse co-occurrence factor and calculated information content, an intelligent and automatic classification of learning objects can be achieved by tagging it as positive or negative for a particular domain.

Keywords: learning object; metadata; taxonomy; co-occurrence; information content;

1. INTRODUCTION

A Learning Object is an Semantically Meaningful Unit (SMU) [1] that is "self-contained" [2]and be able to accomplish its learning objective. Thus, one of the important part of a learning object is "intended to be used for pedagogical purposes"[3]. In order for a piece of content to be considered a learning object, the content must teach something. If the content is not for instructional purposes then it is not a learning object. This distinction is made because not all digital files are learning objects since sometimes their contents are not intended for learning. Still the question remains that how much instructional content a Learning Object has. To answer this, we need to have a vocabulary of terms that defines concepts in a specified domain and the frequency of these terms within a Learning Object. This information about the amount of instructional content in a Learning Object can be stored in its metadata.

According to Feldstein[4], "*Usability in e-Learning is defined by the ability of a learning object to support a very particular concrete cognitive goal.*" The specific sense of the term "usability" suggests particular goals like the context of the evaluation and its pedagogical or instructional intention. For possible context of use and for the evaluation to be feasible, the cognitive goal and its characterization must be described through metadata. Learning Object Metadata include pedagogical attributes such as; teaching or interaction style, grade level, mastery level, and prerequisites[5], however there is a need to add semantics of Learning Object that in return can be used to measure relatedness between two Learning Objects.

To store, search and retrieve Learning Objects, Learning Object Repositories are used. We can say that a Learning Object Repository – LOR is a searchable database that houses digital resources and/or metadata that can be reused to mediate learning. A key process of such repositories is the efficient searching and accurate retrieval of Learning Objects. The question that such LORs fail to answer is; if the resultant Learning Objects are semantically related to the query or not and if they are related, they are related to what degree. Apart from that, queries are based on keywords rather than a Learning Object itself. If a user has a Learning Object, the system needs to find another Learning Object which is semantically similar or opposite.

A survey and report conducted by Reuters found that office-based managers suffered from 'Information Fatigue Syndrome', caused by the frustration of sifting through large quantities of search results[6]. If the user is inundated with large amounts of information, he/she would either waste time manually searching through the results or refining their search queries. So, there is a need to go through the existing technologies and see in the field of information

retrieval, what are the weak areas that need to be improved. What should be an appropriate architecture for a semantic search?

2. LERANING OBJECT METADATA

A number of Learning Object Metadata Standards exist such as Dublin Core[7], IEEE LOM [5] and SCORM [8] etc. These standards focus on the minimal set of attributes needed to allow Learning Objects to be managed, located, and evaluated. However, in this modern world, the pedagogical attribute of an instructional content is far more important. If we evaluate IEEE LOM [5] standards we see there are more than 80 attributes divided in 9 different classes. One of the defined category in IEEE LOM is classification using a predefined taxon path. Normally, author or annotator enters metadata values and assign a taxonomy to a learning object in order to classify it. However, if we can automatically find information content in a particular learning object, we can automatically classify it. We have proposed a system that can automatically classify a learning object and is described in next section.

3. LEARNING OBJECT CLASSIFICATION SYSTEM

Learning Object classification system takes Microsoft PowerPoint® based lectures as an input and classifies it in positive or negative class based on domain terms vocabulary. Various components of this system are term extractor, co-occurrence matrix generator, inverse co-occurrence frequency ICF calculator for each term, Information Content IC calculator for each learning object and classifier. Figure 1 shows each component and their relations. Each of the component is explained in next sections.

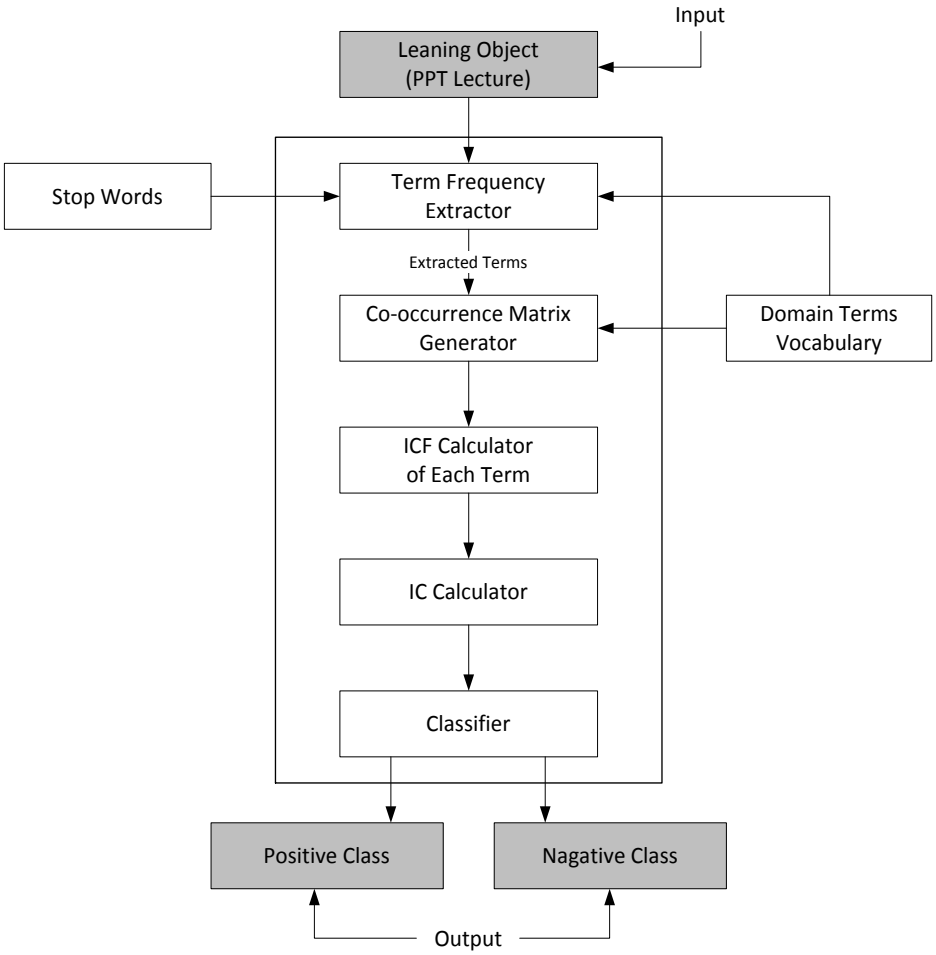

Figure. 1 Proposed framework for learning object classification

4.1. Input data

Learning Objects are the building blocks of any Learning Management System. An LMS is an environment where developers can create, store, reuse, manage and deliver learning content from a central data repository. The LMS generally works with content that is based on a Learning Object model. While no standard definition of a Learning Object exists, a Learning Object generally is referred to a reusable unit of learning. A Learning Object in practice may be a piece of text, sound, an image, a video clip, a flash animation, a Java applet, a web page or an executable program. In our system, we have used only one specific format of learning object that is PowerPoint based lectures delivered at a higher education level. We surveyed different universities of Islamabad, Pakistan and asked different teachers to provide us their set of lectures in one of the following 5 domains of Computer Science. These are;

1. Programming (C++ Programing, OOP, Java, Data Structures etc.)
2. Databases (DBMS, Oracle, SQL Server, PL/SQL etc.)
3. Networking (Protocols, Topologies, LAN, Wireless, AdHoc Networks etc.)
4. Operating Systems (Windows, Linux etc.)
5. Software Engineering (Testing, UML, Fault Tolerance etc.)

Following graph (Figure 2) represents the statistics of data collected from 5 different universities in each domain. A total number of 1050 lectures were collected. Distribution each domain is shown below;

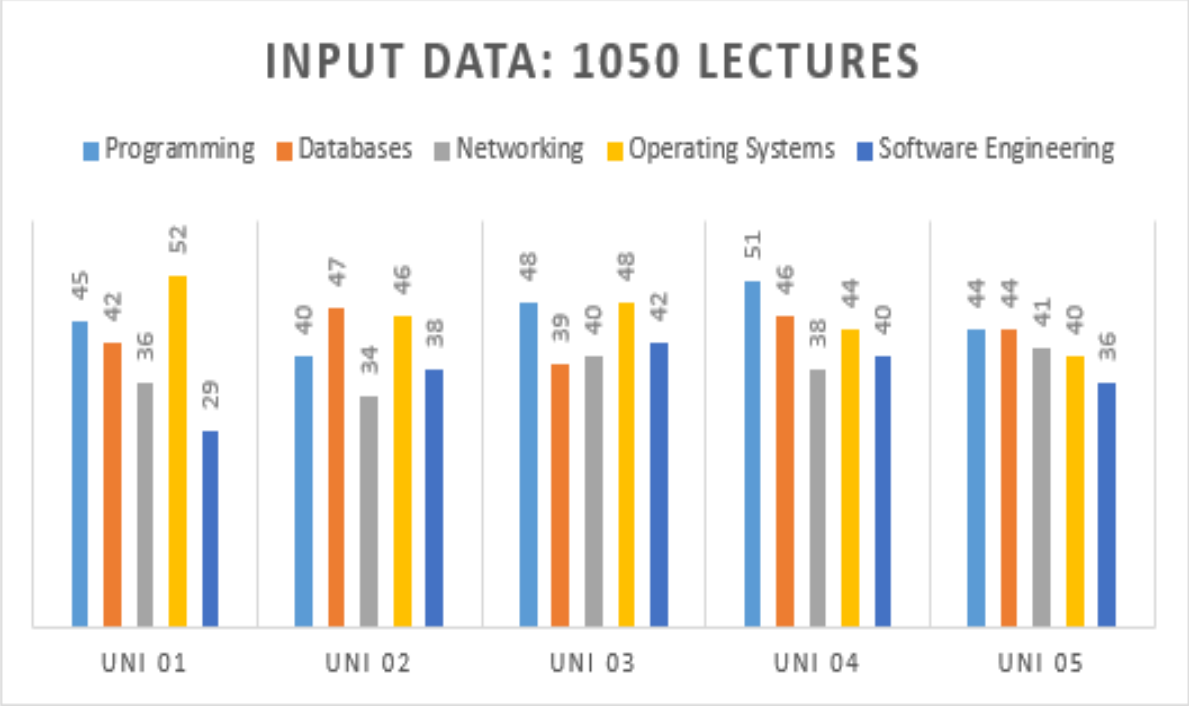

Figure. 2 Input data set

4.2. Domain term vocabulary

Domain Term Vocabulary contains the list of Keywords entered by different experts in a particular domain. We requested experts in each domain to outline keywords that relate to their particular domain and their semantics can mark them as an integral part of the domain. Collection of suck keywords thus formed the Domain Term Vocabulary. However, we do not consider this vocabulary as a closed one, rather it is open in nature, new keywords can be added or old ones can be deleted. Our first collection of each domain vocabulary and the number of keywords in each domain is shown in Figure 3.

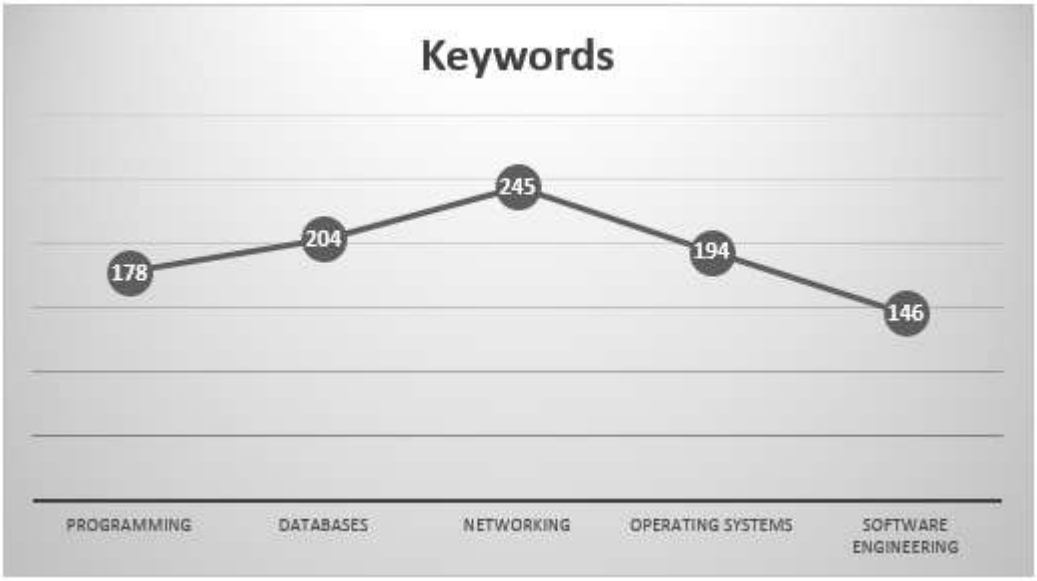

Figure. 3 Domain vocabulary keywords

4.3. Term frequency extractor

Content in a Learning Object is a group of material combined together to teach a single concept. Content material for online courses/tutorials can be any of the following: Explanations, instructions, definitions, images, animations, programs, quizzes, etc. In our case, we are using PowerPoint lectures as a learning object. This module takes learning object that is PowerPoint Lectures in PPT or PPTX format as an input and uses Microsoft Office Interop.PowerPoint Library to load the file. Once file is loaded, it uses two set of vocabularies "Stop Words" and "Domain Terms" to find and extract each term. It uses four steps approach as described below:

1. PowerPoint lectures follow a distinct schema, where each file consists of slides and each slide has header, object and footer area. In the first step, it breaks down loaded PPT or PPTX file in slides, header, object area and footer area in each slide.
2. Removes stop words and filter remaining terms using domain term vocabulary and their occurrence of term in particular area within a slide.
3. Any term appearing in header is more significant as compared to term that appears in object area. Based on occurrence of terms in particular area within a slide, we assign weight and calculate frequency of each term in a particular slide. Weights are assigned on following basis.
    a. Highest weight for term appearing in header of a slide
    b. Medium weight for term appearing in object area of slide
    c. Lowest weight for term appearing in footer of slide

Using these weights, we can calculate term frequency "*tf-idf*" [9]within a slide and is shown in equation 1;

$$tf(i) = tfh(i) \times weight\ of\ header + tfo(i) \times weight\ of\ object\ area + tff(i) \times weight\ of\ footer \quad (1)$$

Where
- tf(i) = term frequency of term "i"
- tfh(i) = term frequency of term "i" in header
- tfo(i) = term frequency of term "i" in object area
- tff(i) = term frequency of term "i" in footer

4. In the last step, term frequency in all slides are combined to find a term frequency within a learning object using equation 2:

$$tf(i)_j = \sum tf(i) \text{ in a Slide}(j) \quad (2)$$

After calculating term frequency of each term, co-occurrence matrix [10] is calculated.

4.4. Co-occurrence matrix

Each learning object has a "bag of words" that is a list of terms within that object. Apart from extracting and calculating frequency of terms within a learning object, a two-dimensional matrix can be created known as "Co-occurrence Matrix"[10]. A sample matrix is shown in Table I.

Table. I Co-occurrence matrix

		Terms									
		T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**
Learning Objects	**LO 1**	3	4	5	0	9	2	0	0	2	1
	LO 2	2	2	2	2	0	5	5	0	2	3
	LO 3	4	3	5	6	0	0	0	1	1	2
	LO 4	2	2	3	2	4	0	0	0	0	0

Each tuple is unique for a learning object. A non-zero term positive frequency describes weighted frequency of each term and its co-occurred terms within a learning object. Based on these co-occurred terms, we can calculate inverse co-occurrence frequency of each term.

4.5. Inverse co-occurrence frequency – ICF

Number of terms in each document that have non-zero positive values describe the first order co-occurrences of keywords. Thus, using this matrix, inverse co-occurrence frequency (ICF) [11]can be calculated with a slightly modified formula and is defined below in equation 3:

$$ICF(i) = \log \left\lceil \frac{\text{Overall number of terms}}{\text{Total number of terms co-occurring with term i}} \right\rceil \times tf(i) \quad (3)$$

Where overall number of terms means the total number of terms that exist in a learning object irrespective of their individual frequency, whereas Total number of terms co-occurring with term "i" means unique set of terms that exist within a learning object. Using the ICF for each term, information content of a learning object can be calculated.

4.6. Information content

To measure specificity for a concept, Information Content (IC) [12]is calculated. If the value is higher, it means the concept is more specific and if value is lower than we can say the concept is more general. Information content is calculated using frequency of concept or terms within a document. Using the ICF value calculated above, we can calculate the IC of each term in a learning object using equation 4:

$$IC(i) = -\log ICF(i) \quad (4)$$

And afterwards, IC of complete learning object can be calculated by summing all the IC for each term as shown in equation 5.

$$IC(\text{learning object}) = \sum ICF(i) \quad (5)$$

4.7. Classification

Based on the Information Content calculated using the domain terms vocabulary, a simple decision can be made. If the IC value is higher, we can assign a learning object to a positive class but if the values are lower, it can be assigned to negative class. A threshold value can be used to form a decision for placing a learning object in relevant class. If a learning object is placed in positive class, we can say that the particular learning object belongs to that

particular domain. If it's placed in negative class then it has to be rechecked for different domain unless it is placed in a positive class for a particular domain. Following flowchart (Figure 4) explains the procedure.

```
Input Learning Object as PPT
Convert PPT to String of Text
Remove Stop Words
Extract Terms and Their Frequencies using Domain Terms Vocabulary
Generate Co-occurrence Matrix using Extracted Terms
Calculate ICF of each Term
Calculate IC of the Learning Object
While the Learning Object has not marked as Positive for a Domain
        Check IC Value with the Domain Value
        If IC value falls in the Domain
                Mark the Learning Object as Positive
        Else
                Mark the Learning Object as Negative
```

Figure. 4 Classification algorithm

4. PROOF OF CONCEPT

A small application is created that has a domain vocabulary. In our case, 1050 selected PowerPoint lectures from different universities and professors was adopted to filter and place them in positive or negative classes for 5 different domains. These 1050 lectures were automatically processed in our application to check the results. Two of the processed lecture screenshot are shown in Figure 5 & 6, one for positive class and one for negative class respectively.

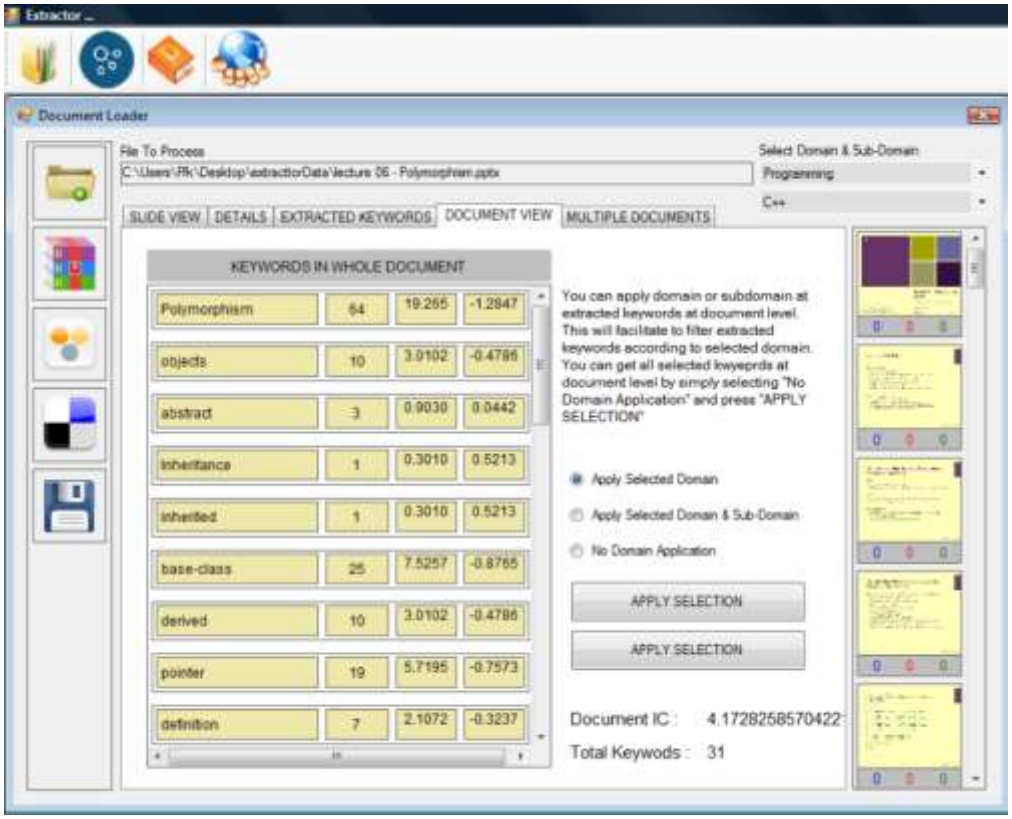

Figure. 5 "Positive" class file

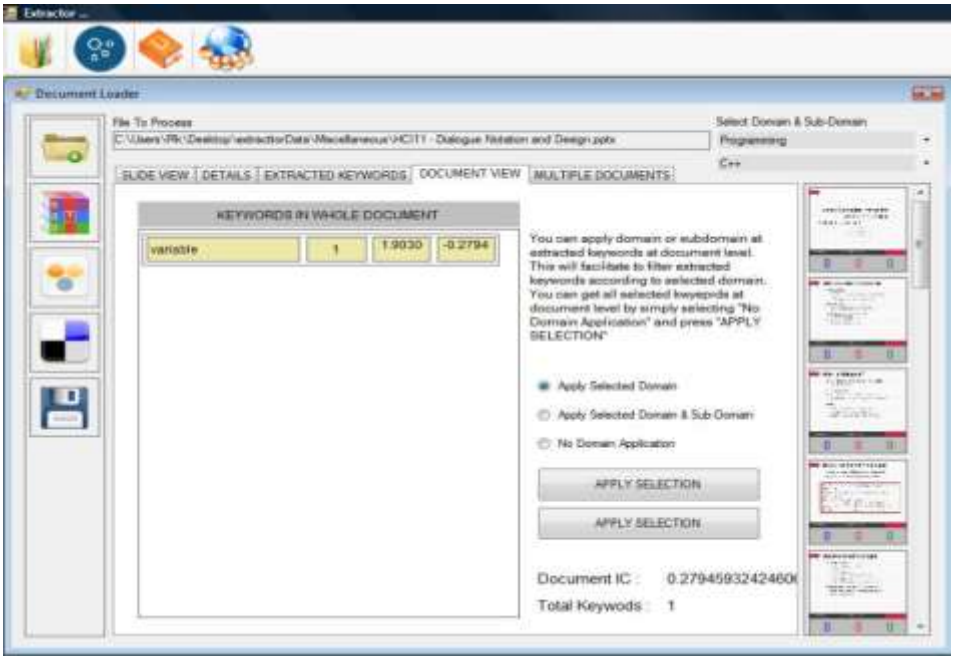

Figure. 6 "Negative" class file

The IC calculated for the selected lectures showed results 0 to 8.07. After careful analysis we marked IC = 2.0 as our threshold value, assigning IC >= 2.0 as "positive" class and IC < 2.0 as "negative" class. Based on this threshold value, results were tabulated and are shown in the figure below.

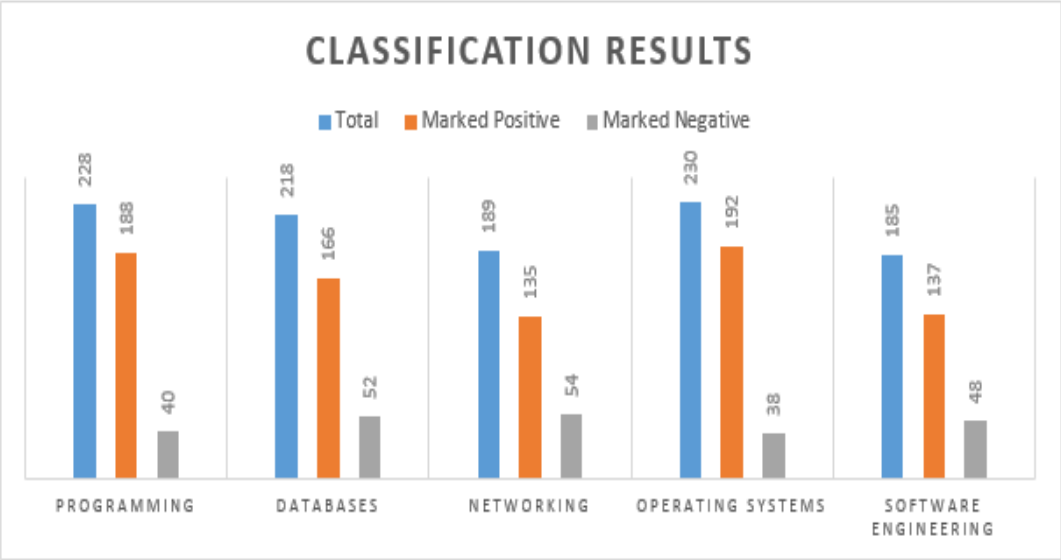

Figure. 7 Classification results

5. RESULTS

Out of possible 1050 "positive" class lectures for all domains, application marked 818 as "positive" for their respective domain and 233 as "negative" or marked them wrongly, giving an accuracy of 77.9%.

6. CONCLUSION

This paper classifies a learning object as positive or negative for a particular domain using Information Content. Information content is calculated using intra terms frequency and their inverse co-occurrence factor ICF. The approach was tested with manual classification and 77.9% accuracy achieved in 1050 lectures.

ACKNOWLEDGEMENT

Special thanks to Faculty of Computer Science in Air University and Capital University of Science and Technology, Pakistan for providing support to complete this research. Moreover, specials thanks to all who have contributed.

REFERENCES

1. Ihsan, I., et al. *Semantically Meaningful Unit-SMU; An Openly Reusable Learning Object for UREKA Learning-Object Taxonomy & Repository Architecture-ULTRA*. in *Computer Systems and Applications, 2006. IEEE International Conference on.* 2006: IEEE.
2. Kevin, O., *An Objective View of Learning Objects.* American Society for Training and Development, 2002. **56**(5): p. 103 - 105.
3. Hesemeir, S.a., *The Tao of Learning Objects: Part One Nature*.
4. Feldstein, *What Is "Usable" e-Learning?* ACM eLearn Magazine, 2002.
5. IEEE (2005) *IEEE Standards for Learning Object Metadata (1484.12.1)*.
6. Wurman, R.S., *Information anxiety. What to do when information doesn't tell you what you*. 1990: New York: Bantam Books.
7. Community, T.M., *Dublin Core Metadata Standards*. 2017.
8. SCORM, *The Shareable Content Object Reference Model*. 2017.
9. Ramos, J. *Using TF-IDF to Determine Word Relevance in Document Queries*. in *Proceedings of the first instructional conference on machine learning*. 2003.
10. Manning, et al., *CS224n: Natural Language Processing with Deep Learning*. 2017.
11. Diederich, Jörg, and W.-T. Balke, *The semantic growbag algorithm: Automatically deriving categorization systems.* Research and Advanced Technology for Digital Libraries, 2007: p. 1 - 13.
12. Pedersen, T. *Information content measures of semantic similarity perform better without sense-tagged text*. in *11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*. 2010.

AUTHORS PROFILE

Mr. Imran Ihsan is currently working as Assistant Professor in the Department of Computer Science at Air University Islamabad, Pakistan. He is also a PhD Candidate at Faculty of Engineering and Computer Science, Capital University of Science and Technology, Islamabad, Pakistan. Mr. Ihsan has more than 20 years of teaching, research & industrial experience. Mr. Ihsan's current research activities are in the field of Semantic Computing, Ontology & Knowledge Engineering, E-Learning and Human Computer Interaction.

Mr. Faisal Fayyaz Kiyani is currently working as Lecturer in Department of Computer Science at Air University Islamabad, Pakistan. He is also a PhD Scholar at Faculty of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan. Mr. Faisal has more than 12 years of software design & development experience in the industry and over 4 years of teaching experience. His areas of interest include Semantic Web, Web Mining, IR, NLP and Distributed Systems.